A new scenario for the early evolution of Mycobacterium tuberculosis

Yann Blouin

UNIVERSITE PARIS-SUD

ÉCOLE DOCTORALE : *Gènes, Génomes, Cellules*

Institut de Génétique et Microbiologie

Génomique

THÈSE DE DOCTORAT

Soutenue le 04 septembre 2014

par

Yann BLOUIN

A new scenario for the early evolution of *Mycobacterium tuberculosis*.

Directeur de thèse :	Gilles VERGNAUD	DGA, Bagneux

Composition du jury :

Rapporteurs :	Roland BROSCH	Institut Pasteur, Paris
	Noel H. SMITH	AHVLA, Addlestone (UK)
Examinateurs :	Vincent JARLIER	CNR Tuberculose, Paris
	Guislaine REFRÉGIER	Université Paris-Sud, Orsay
Président du Jury :	Pierre CAPY	Université Paris-Sud, Gif sur Yvette

Acknowledgments.

A work such as this thesis is not something that you can manage to realize on your own. It is the result of the contributions of many different people, and I truly want to thank them all for their help, direct or indirect.

First let me express all my gratitude to Dr Gilles Vergnaud and Dr Christine Pourcel for the part of my life that I have spent in their team. Words are a poor medium to express my thanks to these exceptional researchers for all that they have given me during these years. If as the saying goes, "time is money", then it is sure that I owe them a fortune for the time they have invested in my scientific education.

I would like to thank the members of the jury for their presence and the fact that they accepted to take part in the final step of this work. I am especially grateful to the two referees Pr Noel H. Smith and Pr Roland Brosch for their reading of this manuscript and the improvements that they suggested. Thanks also to Dr Refrégier, Pr Jarlier and Pr Capy for examining my work, and to Dr Derzelle and Pr Lespinet for taking part in my thesis committee during those three years.

This work has taken three years and I have spent all these years in a great team. My thanks go to all the GPMS team members that have worked and discussed with me during this time: Yolande and Philippe, the wise ones, Christiane and Daniel, who finished their PhD while I was doing mine, Christophe for the great discussions that we shared, Libera and Cedric who supported me during these last months of redaction, and all the people that have stayed in our lab during this period: Rim, Julien, Jérémy, Adrien, Maxime, Nicolas. And Guillaume, who should be awarded an honorary team membership, even if he stayed here only two weeks. I have also to express my thanks to our "neighbors" of the IGEPE team, for the lunch breaks that we spent together, and to the other members of the second floor.

Of course everything was made possible because of the great infrastructure of the IGM and the GGC doctoral school, so I also want to thank all the people that belong to these two places, especially Murielle, Marie-Claude, Catherine, Aristide, Mazyar and Thierry. Many thanks also to the DGA for allowing me to spend these last three years studying in this field to gain the expertise needed for my future work.

And last but not least, I want to mention my parents for their unswerving support. I know that they have always believed in me, and their help has been especially required and welcome during this work.

Content

I - INTRODUCTION **8**

1) Theoretical aspects: biological and ethnological considerations. 8

1.1) Generalities about bacterial populations and bacterial genomics. 8

1.2) About *M. tuberculosis*: the discovery of the Koch bacillus. 11

1.3) The biology of *M. tuberculosis*. 14

1.4) The *M. tuberculosis* complex and molecular epidemiology. 17

1.5) The MTBC phylogeny. 22

1.6) Horizontal gene transfer and mutation rate. 28

1.7) About "*M. canettii*": discovery and pathogenicity. 29

1.8) Phylogenetic analysis and relationship with the MTBC 33

1.9) Environmental Mycobacteria. 36

1.10) Historical records for tuberculosis. 38

1.11) History of the Horn of Africa. 42

1.12) Summary : state of the art of the knowledge at the start of this work. 44

2) Technical aspects: sequencing and bio-informatics. 46

2.1) Specificities of high-throughput sequencing data analysis. 46

2.2) Computing challenges . 52

II - RESULTS **56**

3) Description of the results obtained during this work and of their background. 56

3.1) Preliminary considerations. 56

3.2) Article I: Significance of the Identification in the Horn of Africa of an Exceptionally Deep Branching Mycobacterium tuberculosis Clade. 61

3.3) Article II: Progenitor “Mycobacterium canettii” Clone Responsible for Lymph Node Tuberculosis Epidemic, Djibouti. 62

III - DISCUSSION **63**

4) Emergence and evolution of the MTBC. 63

4.1) Model of emergence. 63

4.2) Dating the model. 73

5) Conclusion: about the use of SNPs for phylogenetic purposes and the perspectives of high-throughput sequencing. 84

IV - BIBLIOGRAPHY **88**

V - ANNEXES **96**

A) Methods. 96

B) Details of other publications. 104

C) List of scripts. 109

Foreword

Bacterial pathogens remain one of the major causes of death worldwide and more generally of economic losses. Man has no predators, but he has to fight against bacteria in a non-ending conflict as pathogens keep evolving to outsmart our defenses. In this regard understanding the emergence and evolution of bacterial pathogens is essential in order to better control outbreaks, or even to prevent the appearance of new ones that could pose a threat to the global human population.

To determine the factors that are important in the study of the evolution of pathogenic bacteria, I decided to use inductive reasoning focusing on a particular pathogen. My subject of study is *Mycobacterium tuberculosis*, a major worldwide health problem, and a widely studied bacterium. Consequently, data are continuously being produced, and are available publicly for incorporating into any global analysis. By trying to understand how this bacillus emerged and evolved into its extant lineages, I hoped to be able to infer global traits that shape bacterial evolution and host-pathogen interactions. For this purpose, I also took advantage of *M. tuberculosis* closest neighbor, the environmental opportunistic pathogen "*Mycobacterium canettii*". These two bacterial population structures will prove to be highly complementary.

List of abbreviations used in this work

aDNA	ancient DNA
BCE	**b**efore **c**urrent **e**ra
bp	**b**ase **p**air
CAS	**C**entral **As**ia, a.k.a. Lineage 3
CDS	**C**oding **D**NA **S**equence
CRISPR	**C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats
DNA	**D**esoxyribo-**N**ucleic **A**cid
EAI	**E**ast **A**frica & **I**ndia, a.k.a. Lineage 1
GUI	**G**raphical **U**ser **I**nterface
HIV	**H**uman **i**mmunodeficiency **v**irus
ICDS	**I**nterrupted **Co**d**i**ng **S**equence
IS	**I**nsertion **S**equence
kY	thousand **y**ears
LAM	**L**atin **Am**erica
MDR	**m**ulti **d**rug-**r**esistant
MLSA	**M**ultiple **L**oci **S**equence **A**nalysis
MLST	**M**ulti**l**ocus **S**equence **T**yping
MLVA	**M**ultiple **L**oci **V**NTR **A**nalysis
MRCA	**M**ost **r**ecent **c**ommon **a**ncestor
MST	**M**inimum **S**panning **T**ree
MTBC	**M**ycobacterium **t**u**b**erculosis **c**omplex (or M. tuberculosis complex)
NDT	**N**eolithic **D**emographic **T**ransition
NGS	**N**ext **G**eneration **S**equencing
RFLP	**R**estriction **F**ragment **L**ength **P**olymorphism

SNP	**S**ingle **N**ucleotide **P**olymorphism
RD	**R**egion of **D**eletion
TB	**T**u**b**erculosis
VNTR	**V**ariable **N**umber of **T**andem **R**epeats
XDR	e**x**tremely **d**rug-**r**esistant

Tables and figures

Figure 1: Ecotype model of evolution. 10

Figure 2: Major figures in TB research..... 13

Figure 3 : M. tuberculosis life cycle..... 15

Figure 4 : Schematic principle of major genotyping techniques..... 18

Figure 5: Examples of raw typing results..... 21

Figure 6: Phylogenetic tree based on deletions..... 24

Figure 7: Spoligotypes of the different lineages and subgroups of the MTBC. 25

Figure 8 : Morphology of tuberculosis-causing colonies. 29

Figure 9: Map of the Horn of Africa..... 30

Figure 10: "Mycobacterium canettii" diversity..... 36

Figure 11 : Global tree of Mycobacteria..... 37

Figure 12: Map of the evidence for historical occurrences of tuberculosis during human history. 42

Figure 13: Minimum spanning tree of the MTBC lineages. 46

Figure 14: Major figures in DNA research. 48

Figure 15: Example of reads mapping on a bacterial genome..... 51

Figure 16: Details of the four ICDSs differentiating "M. canettii" strains from the MTBC members. 59

Figure 17: Evaluation of the interrupted status of the 81 ICDSs on several strains from our collection. 60

Figure 18: Diagram of the proposed population structure and evolutionary history of "M. canettii" and the MTBC..... 68

Figure 19: Hypotheses regarding extant "ancestral" lineages of M. tuberculosis before the Djibouti study. 70

Figure 20: Phylogenetic tree of the superlineages of the MTBC. 71

Figure 21: Linear model of evolution for the M. tuberculosis complex. 73

Figure 22: Ratio of non-synonymous to synonymous variations within the MTBC..... 77

Figure 23: Probable location of the ancient Land of Punt. 81

I - Introduction

1) Theoretical aspects: biological and ethnological considerations.

1.1) Generalities about bacterial population and bacterial genomics.

Since the early work of Linnaeus, the notion of species has been hotly debated in all the fields of biology. Until the advent of modern genomic techniques, the attribution to one species or another was based on the observation of shared phenotypic characteristics, and the capacity to reproduce with other members of the same species. These characteristics constituted also the basis to reconstruct the "Tree of Life", as described by Charles Darwin (Darwin, 1859).

If the definition of species has sometimes proven difficult in the field of "macro-biology" and sexual reproduction, it is even harder when one considers microbes. The determination of shared traits is more complex, and it is more difficult to infer the relationships between different bacterial strains. This problem has fuelled theoretical work based upon the very idea of bacterial species. A practical solution has been proposed with the hybridization of DNA from different strains, which is a proxy for the genetic distance between them. It is used to determine if two strains are closely related enough to belong to the same species. A threshold of 70% has been empirically adopted in order to fit reasonably with previous definitions of species. However such a definition is not likely to have a robust biological meaning. One theoretical solution to this species definition problem is presented by Cohan (Cohan, 2001). His work further developed the notion of "ecotype", first introduced by Turesson in 1922 (Turesson, 1922; Turrill, 1946) in the field of vegetal ecology. An ecotype would be a group of cells occupying the same ecological niche and sharing phenotypic and ecological properties, reflected by homogenous DNA characteristics. An ecotype would share a number of properties with the definition of species for eukaryotes but at the same time be more appropriate for prokaryotes (Cohan, 2001). Inside an ecotype, the genetic diversity has to be under some sort of constraints caused by cohesive forces (essentially selection).

In this view, one key element, specific for bacteria, is the way recombination occurs in these organisms, and the impact it has on shaping their genetic content, driving evolution and allowing the emergence of new ecotypes. In bacteria, genetic exchanges can happen relatively easily when the proper conditions are present. When there is no recombination (i.e. no uptake of outside genetic material) in a bacterial population, this population is called "clonal". Each cell's genome is the direct vertical evolution from its ancestors. This gives rise to perfect tree-like phylogenies. Recombination alters the "vertical" phylogenetic signal, and one then needs to envision the population as a network instead of a linear tree (Smith *et al.*, 1993). When exchanges are frequent, the population is called "pan-mictic".

Besides recombination, other mechanisms impact the evolution of an ecotype, as presented by Cohan. Genetic drift and selective sweeps are two other major elements that contribute to shape the genetic diversity inside a given ecotype. Both are linked to the appearance of random mutations during the replication of any bacterium. These random mutations can be deleterious for the individual but they can also provide a selective advantage. If selection takes place, given enough time, these mutations can replace the anterior base-state in all the population (by outcompeting the older genotype). This is a "selective sweep". A selective sweep can happen at any time; two genotypes can coexist during a long period until a change in the environment is advantageous to one of these genotypes that can then outcompete the other one (Figure 1). This aspect is one characteristic of the definition of an ecotype: the "adaptive mutant" will outcompete all the other members of its ecotype but not the strains belonging to a different one. Neutral mutations will also eventually be fixed by genetic drift, in the absence of selection. In this case the diversity observed is only governed by random sampling of the population.

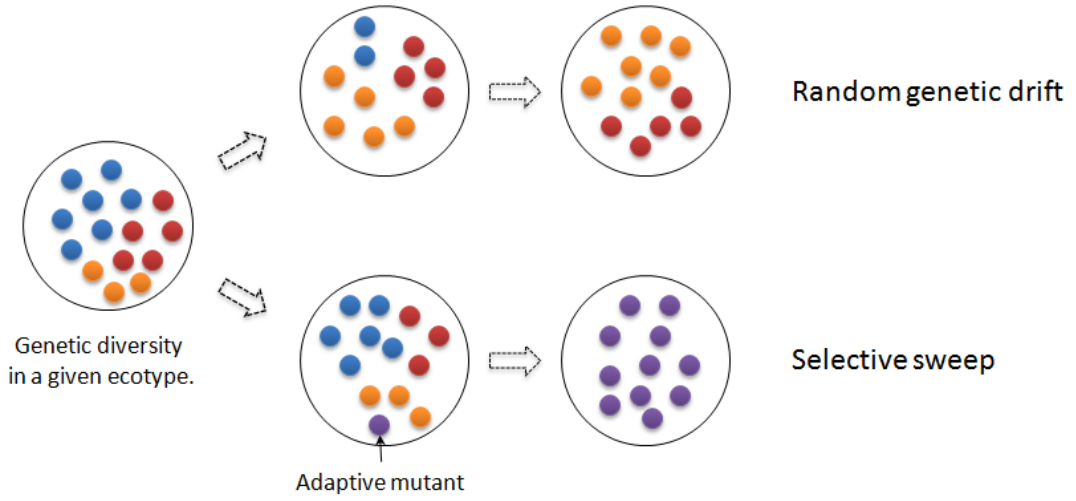


Figure 1: Ecotype model of evolution.

This diagram represents the two main evolution routes in the ecotype model as developed by Cohan. Initially the ecotype is composed of diverse genotypes (shown with colors); in the random genetic drift one is lost by stochastic processes whereas in the selective sweep, a new genotype appears and outcompetes the others.

The effects of selection are of different kinds, and the way selection is acting on bacterial genetic diversity is still a matter of discussion. One way to assess the effects of selection, now that it is possible to have access to the genetic sequence of a given organism, is to consider the ratio of non-synonymous to synonymous substitutions, called dN/dS ratio. To be able to consider this ratio, the single nucleotide variants (which can be identified as Single Nucleotide Polymorphisms or SNPs) between two genomes have to be determined first. Substitutions appear at random positions along the genome due to errors during the replication process. If there is no selection of the produced genetic sequences, and not taking into account the potential mutator effect of specific secondary DNA structures, all sites have the same probability to be mutated and to be conserved in the final set. In that case the dN/dS ratio is equal to 1.

Two aspects of selection can modify the observed dN/dS ratio, in one direction or the other. First there is purifying selection. If its population is "locked" in an equilibrium configuration, deleterious mutations will be counter-selected, and the non-synonymous substitutions frequency will diminish (negative selection). The dN/dS ratio will be lower than 1 (the smaller it is the higher the cumulative effect of purifying pressure). The opposite scenario happens when the environmental conditions change

brutally from an equilibrium state to a new one. The ancestral genotypes are not adapted to the new conditions, and therefore the mutants bearing mutations conferring a selective advantage are selected. This may result in an increase in the observed frequency of non-synonymous mutations, and a dN/dS ratio greater than 1, at least in parts of the bacterial genome under positive selection (pressure for change) (Kryazhimskiy *et al.*, 2008).

As mentioned before, bacteria may be subjected to horizontal gene transfer events, during which they receive genetic material from another ecotype, possibly more distantly related. These events can be identified when looking at the dN/dS variations along the genome, as the history of the selective pressures applied on the transferred genetic regions will differ from that of the rest of the genome. As this depends on multiple factors, there is no way to predict how this ratio evolves throughout time.

An important remark concerning deleterious mutations is that they might not be counter-selected. *In silico* simulations have established that deleterious mutations could open evolutionary pathways that will prove successful after several steps of additional mutations, leading to a new equilibrium that would otherwise not have been accessible. This is why determining the effects of selection on non-synonymous substitutions is not always straightforward. Even if they are potentially slightly deleterious, they could prove beneficial in combination with some posterior modification of the genome (Covert *et al.*, 2013).

In order to address some of these issues, especially the relationships between ecotypes and the mechanisms governing selective pressure and the emergence of new variants, we have chosen *M. tuberculosis* as a model for reasons presented in the following pages.

1.2) About M. tuberculosis: the discovery of the Koch bacillus.

M. tuberculosis is the causative agent of human tuberculosis (TB). It was identified in Germany at the end of the nineteenth century by Robert Koch (Koch, 1882) (Figure 2). Owing to the development of new staining methods, he was the first

to observe the bacillus causing the disease which had plagued Humanity for many centuries. The new staining technique involved two distinct colorations to highlight the bacterium from the background. Koch was awarded the Nobel Prize in 1905 for the discovery of what is still often called "Koch's bacillus". In Europe the tuberculosis epidemic was raging since at least two centuries, profoundly affecting the society, and killing several millions of people. Its traces are visible in the cultural legacy of this period, when consumption (as tuberculosis was then called) affected many people and seemed to result in an unavoidable death. As it was a major cause of concern because of its repercussions on public health, tuberculosis was investigated by major actors in the medical field. The French physician René Laennec described it in his 1819 publication about auscultation (Laennec, 1819). Jean Antoine Villemin, another French physician, established the contagious status of this disease in 1865 (Villemin, 1865). At this time, Villemin was working at the French Military Hospital of Val-de-Grâce in Paris. He established that TB could be transmitted from humans to rabbits, but also from cattle to rabbits or between rabbits. But until Koch's discovery, the exact nature of the causative agent of tuberculosis was unknown, even when many were trying to cure patients of this disease (Daniel, 2006; Cardoso Leão *et al.*, 2007).

Following the discovery of the tubercle bacillus, Koch tried to develop a vaccine that could be used to prevent people from being contaminated by tuberculosis. His efforts were not successful. They led him to the discovery of tuberculin, but this compound did not prove efficient as a vaccine, leading even to some deaths. It was later used as a starter by Clemens von Pirquet, an Austrian scientist, to develop a detection test to identify patients who had been exposed to tuberculosis at some point during their life. Von Pirquet had discovered the ”allergy” phenomenon, and used that knowledge to extend Koch preliminary tests with tuberculin (the "Pirquet Reaction"). The test was then further refined by a French researcher, Charles Mantoux, who developed the "Mantoux test", which is since then considered as a reference for the screening of TB infection.

The efforts to develop a vaccine, similar to what had been done by Jenner with smallpox (Jenner, 1798), were also pursued in France by Calmette and Guérin, working

at the Pasteur Institute in Lille. They tried to attenuate a strain of *Mycobacterium bovis*, a species responsible for a tuberculosis-like disease in cattle and sometimes in humans. At first their task seemed impossible, as *M. bovis* proved to be as pathogenic as *M. tuberculosis* in humans. After several years of efforts they obtained in 1919 an attenuated strain that has thereafter been used to vaccinate patients, especially children during the twentieth century. The first human vaccine trial took place in 1921. This new vaccine was called BCG vaccine, for "vaccin Bilié de Calmette et Guérin" (but the corresponding strain is also called BCG, for "Bacille de Calmette et Guérin"). BCG is now widely used as a mean of vaccination in children, even if its efficiency against pulmonary tuberculosis in adult populations, in regions where TB is endemic, is questioned (Martín *et al.*, 2007).

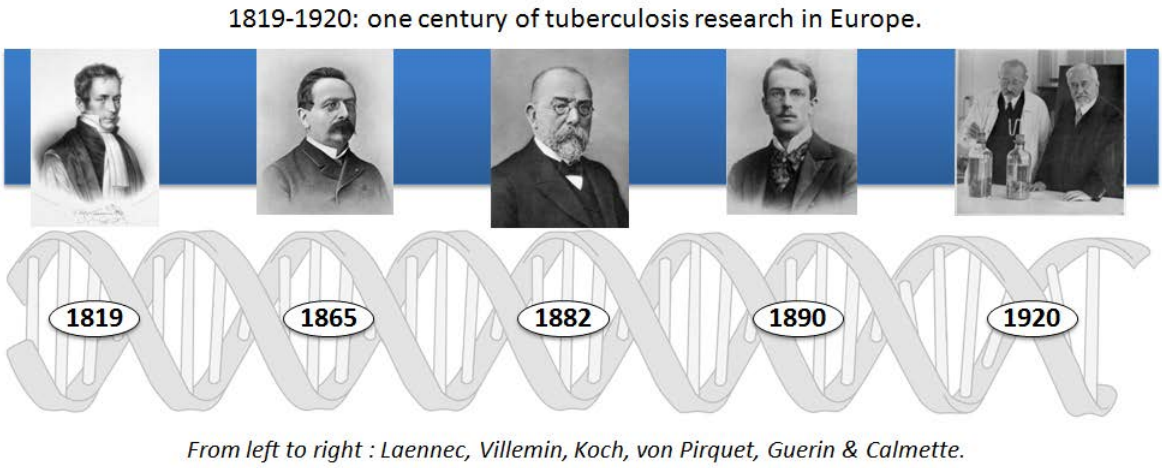


Figure 2: Major figures in TB research.

Here are presented the portraits of five major figures in TB research, with the date of their key contribution.

The next breakthrough in the fight against tuberculosis was the discovery of the first antibiotics efficient against *M. tuberculosis*, namely streptomycin in 1944, and isoniazid in 1952. However, already in the first therapeutic uses, appeared one major characteristic of the struggle between the bacteria and human research: the emergence of resistant bacterial strains selected by the antibiotic. Other antibiotics were found that proved effective against *M. tuberculosis*, and multi-antibiotics chemo-

therapy was established as the most effective way to cure TB patients. This led to the closure of the sanatoria that had been the hallmark of nearly a century, since their appearance in the second half of the nineteenth century. It illustrates the major change that anti-mycobacterial drugs brought to the field of TB treatment.

Tuberculosis has almost disappeared in developed countries, however the tuberculosis epidemic has not receded to the same extent in developing countries where it kills yearly between one and two millions people (WHO, 2013). It seems to be strongly linked to poverty and malnutrition. The emergence of antibiotics-resistant strains (multi-drug resistant, MDR or extremely drug resistant, XDR) leads some to fear that tuberculosis could become deadly again even in developed countries. Moreover, according to the World Health Organization, tuberculosis is the first cause of death for immuno-depressed patients, and is therefore a major health issue in countries where the HIV-pandemic is raging, like in many African countries. Today the struggle against tuberculosis is far from over.

1.3) The biology of M. tuberculosis.

M. tuberculosis is an obligate intra-cellular pathogen, which means that it spends most of its life cycle inside its host's cells. It belongs to the genus *Mycobacterium*, bacteria that share a remarkable composition of their cell wall, containing mycolic acids. Their DNA is characterized by a high G+C content (between 61 and 71%) (Barrera, 2007). This genus encompasses some other highly dangerous human pathogens, such as *Mycobacterium ulcerans* and *Mycobacterium leprae* (cf. section 1.9). On a conceptual level, they could be considered as Gram-positive bacteria, as they do not possess an outer cell wall. However the proteins attached to this cell wall are lipids, and not proteins or polysaccharides. The particular composition of their cell wall classifies them as acid-fast, meaning that these bacteria do not retain the classical dye used for staining. Their cell wall makes *Mycobacteria* resistant to harsh environmental conditions, and also gives them hydrophobic characteristics. It is believed that it may confer a propensity to resist some antibiotics, such as beta-lactamases. It may also be responsible for the slow growth rate displayed by some of

them. For instance, *M. tuberculosis* divides each 12 to 24 hours, compared to the 15 to 60 minutes sufficient for most cultivable bacteria. This growth rate could also be related to the presence of a single operon controlling RNA synthesis, which could explain the ten-fold difference observed in the RNA chain elongation rate between *Escherichia coli* and *M. tuberculosis* (Harshey *et al.*, 1977).

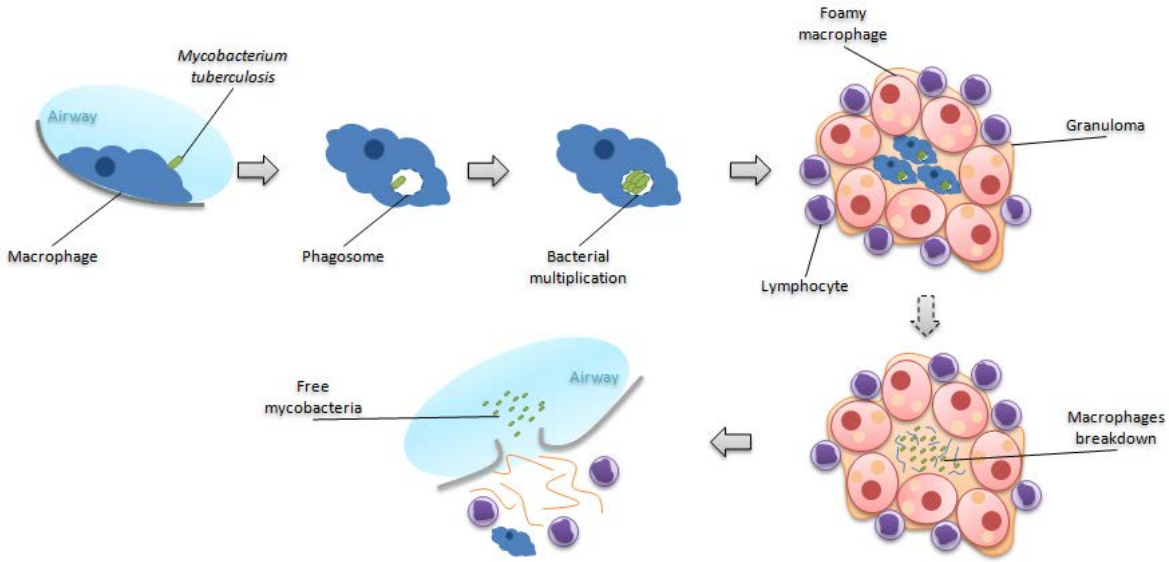


Figure 3 : *M. tuberculosis* life cycle.

This drawing is inspired by (Russell et al., 2010). The infection occurs when *M. tuberculosis* bacilli are (in most cases) inhaled. They are then engulfed in macrophages and multiply in early phagosomes. At the site of infection, other immune cells are recruited, leading to the formation of a granuloma. The death of the infected macrophages leads to the release of multiple free bacilli.

In humans the bacterium takes advantage of the natural defense mechanisms of the individual it invades, by entering mononuclear macrophages (Figure 3). The pathogen has developed specific ways to evade phagocytosis by blocking the phagosome fusion with the lysosome, thus preventing the destruction of the bacterial cell. This impairs the acidification of the phagosome, keeping the environmental conditions safe for the bacteria that can then replicate inside it (Hernández-Pando *et al.*, 2007).

The main localization of an infection by *M. tuberculosis* is at first the upper area of the pulmonary lobe. The infection is usually caused by the inhalation of tiny droplets

containing bacilli. Some studies have established that ten bacterial cells were enough to start the infection in 50% of the cases (Balasubramanian *et al.*, 1994; Dean *et al.*, 2005). Once inside the pulmonary alveoli, the tubercle bacillus causes a natural inflammatory response, leading to the requisition to the site of the infection of cells of the immune system. Macrophages then begin to phagocyte bacterial cells. Two outcomes are possible: either the defense system is strong enough and the macrophage is capable of controlling the bacterial cell division, or the bacterium successfully evades the phagocytosis mechanisms by impairing the fusion of the phagosome with the lysosome in the infected cell. In that case, the infection can spread through the blood stream or the lymphatic system, where the infected macrophage is travelling, possibly transmitting the infection to other locations in the body of the host, such as the kidneys, the brain (tuberculous meningitis) or the bones (osseous tuberculosis, also called Pott's disease when it affects the spine) (Kritski *et al.*, 2007).

If the natural immune response progresses normally (i.e. if the immune system of the host is not deficient) the infected macrophages are surrounded by other immune system cells, leading to the formation of a granuloma, typical of an infection by *M. tuberculosis*. This enables the containment of the pathogen. In that case the infection is controlled unless the immune system of the host is weakened. This is called latent infection. In most cases there are no traces of primary tuberculosis infection (i.e. the patient does not develop symptoms when the organism is invaded for the first time by the pathogen) and the disease can stay latent for the whole life of the individual. It is only when reactivation occurs that the symptoms become apparent. According to current estimates, 5 to 10% of infected people will develop clinical signs of tuberculosis during their life. In that case, pulmonary tuberculosis is the most common manifestation in adults, even if the lymph nodes can also be affected. When the disease is active in immunocompetent individuals (i.e. if the pathogen manages to evade the immune response, and is left untreated) tuberculosis leads to death in 50% of cases, and to chronicity in about 25% of cases. Natural cure occurs for one quarter of the patients with active disease. The reactivation usually takes place three to five

years after the primary infection. In HIV-positive patients, tuberculosis is the major cause of death world-wide because the immune system is too weak to cope with the burden of tuberculosis infection.

Another form of tuberculosis, besides the pulmonary one, is lymph nodes infection (lymphadenitis). This form of TB infection is particularly frequent in children under five, and is often localized in the neck region. However, in this case, it may also be caused by non-tuberculous mycobacteria, particularly in hot weather countries (90% of the cases in young children). Above the age of 12, 90% of lymphadenitis cases are due to *M. tuberculosis* (Johnson *et al.*, 1998).

1.4) The M. tuberculosis complex and molecular epidemiology.

Following the work of the nineteenth century researchers, it was known that *M. bovis,* the animal pathogen also able to cause tuberculosis in humans, was different from *M. tuberculosis*. However the exact phylogenetic relationship between these two species was undetermined. Biochemical tests led to the naming of additional species such as *M. africanum*, and animal-specific families. They were eventually all gathered into the *Mycobacterium tuberculosis* complex (MTBC). The commonly accepted view early in the study of the MTBC was the emergence of *M. tuberculosis* from *M. bovis* as a result of cattle-domestication, and subsequent transmission of the disease from animals to humans. This was later demonstrated to be unlikely. The development of molecular typing tools during the last decade of the twentieth century enabled to unravel the diversity inside the complex, and allowed to establish the current view for the emergence of this human pathogen (cf 1.5).

One of the first molecular markers used to study the MTBC was an insertion sequence (IS). IS are mobile genetic elements that can move inside a genome, and as a result usually present a certain diversity in their number and insertion sites per genome in a bacterial population. This polymorphism can therefore be used to characterize the strains, thus qualifying IS elements as typing markers. For the MTBC, the most informative IS is IS6110, first described by Thierry *et al.* (Thierry *et al.*, 1990),

later becoming an internationally recognized typing tool (van Embden *et al.*, 1993). The typing method, sometimes called DNA fingerprinting, is based on the analysis of IS restriction fragment length polymorphism (RFLP) by hybridization: the studied DNA is digested by a given restriction enzyme, run on an agarose gel, transferred to a membrane (procedure called "Southern blotting") and the membrane is hybridized with a labeled IS6110 probe. The different IS loci are distinguished because the position of the restriction sites flanking the IS element will differ at the different insertion sites. A strain is characterized by its RFLP pattern (Figure 4). The first RFLP investigation of *M. bovis* diversity was published in 1994 and used a collection of 153 strains (van Soolingen *et al.*, 1994). This study identified a previously unknown heterogeneity in *M. bovis*, with genotypes probably reflecting the host preference of the strains, as the strains originated not only from cattle, but also from zoo animals. The endemicity of some RFLP patterns present in the dataset was also established.

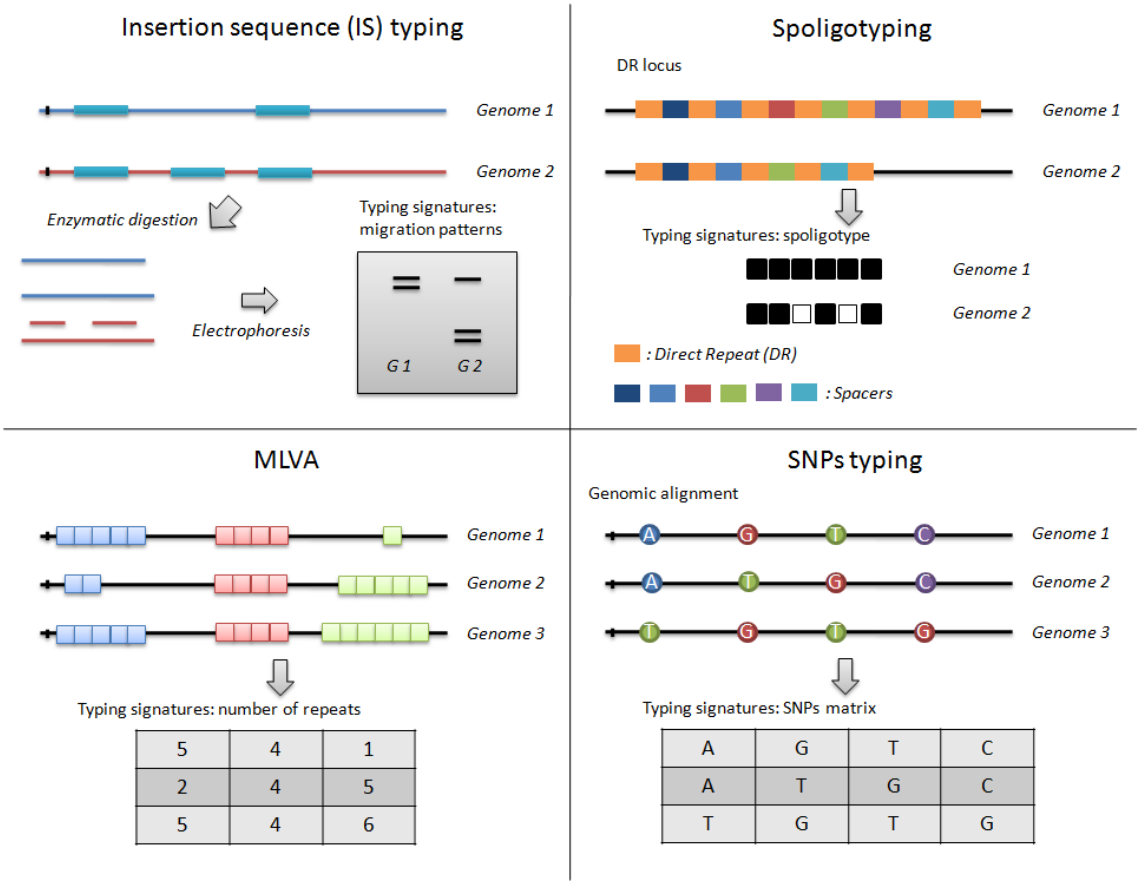


Figure 4 : Schematic principle of major genotyping techniques.

Top left panel: IS are represented as blue bars, and the genomic sequences are in blue and red to illustrate the generation of fragments of different size after enzymatic digestion. Top right panel: the DR region is schematized by the orange bar whereas spacers are color-coded, leading to the determination of a binary code, the spoligotype. Bottom left panel: the different VNTR loci are shown by different colors and for each locus the number of repeats is represented by the number of colored squares. Bottom right panel: the variation of bases at different positions along the genome is emphasized by a colored node.

After some years of use, it appeared that the IS6110 genotyping method suffered limitations. For instance, comparison of results produced in different laboratories was technically demanding. Consequently, owing to the decreasing costs of sequencing technologies, new typing techniques were envisioned such as partial sequencing of the bacterial genome. A study considered the use of 26 structural genes in order to analyze the existing Single Nucleotide Polymorphisms (SNPs) (Figure 4), and infer some evolutionary mechanisms from this information (Sreevatsan *et al.*, 1997). In this study, Musser and col. found that many mutations seemed to be linked with antibiotic resistance. The genes *katG* and *gyrA* showed the highest number of synonymous mutations compared to the other genes considered. Using SNPs not subjected to selection by antibiotic resistance, they defined three genetic groups inside the MTBC, encompassing different known lineages. Then the authors used this subdivision to try to infer a possible evolutionary scenario for the emergence of the *M. tuberculosis* complex, postulating an anteriority of the animal adapted lineages compared to *M. tuberculosis* strains *sensu stricto*. Furthermore they postulated that the limited genetic diversity observed between MTBC strains could be linked to a bottleneck, due to the acquisition of the obligate pathogen way of life when the progenitor of the MTBC became a human pathogen, some 15,000 to 20,000 years ago. However this partial sequencing approach, although providing a couple of phylogenetically robust markers, was of little use for epidemiological purposes as it had a very limited discriminatory power for a very high cost.

At the same time van Embden and coll. from the Netherlands proposed a new typing tool for the genetic characterization of *M. tuberculosis* strains: spacers oligotyping, commonly referred to as spoligotyping (Kamerbeek *et al.*, 1997). This typing scheme is based on the peculiar properties of a specific region of the bacterial

genome, the direct-repeat (DR) locus. It presents a succession of conserved repeats (DRs) separated by variable regions called spacers. This region is a CRISPR locus (for Clustered Regularly Interspaced Short Palindromic Repeats) (Figure 4). In some species CRISPR and CRISPR-associated genes (Cas) play a role in immunity against bacteriophages, by acquiring fragments of the invading genome, which will help protect the bacteria during the next encounter with the phage (Pourcel *et al.*, 2005). In the members of the *M. tuberculosis* complex this locus is no longer acquiring new spacers, but progressively losing them by deletion, and contains one or more IS elements suggesting that it might be inactive. As a results, MTBC strains can be conveniently characterized by the set of spacers they possess, which is necessarily a subset of all the spacers known for the complex. This enabled the development of a method that is simple, robust and rapid, making it one of the core methods of genetic characterization of MTBC strains. The typing of each strain requires a single PCR amplification, and more than forty strains can be typed simultaneously by hybridizing the labeled PCR product on a spoligotyping membrane (Figure 5). The result can be easily coded (Figure 4). Later this technique was automated to study larger numbers of samples (e. g. Luminex technology (Cowan *et al.*, 2004; Abadia *et al.*, 2011)). To date a database of 7,104 profiles, called spoligotypes, can be found in the SITVIT Web database (http://www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE/), grouping 58,187 strains from 102 countries. There is also another major database dedicated to spoligotyping data for the characterization of *M. bovis*, accessible at http://www.mbovis.org. It contains more than 2,000 profiles from *M. bovis* strains.

In 1998 Variable Number of Tandem Repeats (VNTRs) were introduced in the field of TB genetic analysis. Such genetic elements, present at multiple loci in the *M. tuberculosis* genome, are made of tandem repeats that may differ in copy number from strain to strain (Figure 4). The first MLVA (Multiple Loci VNTR Analysis) typing scheme for *M. tuberculosis* comprised five loci selected among eleven loci (Frothingham *et al.*, 1998). The authors determined the diversity at these loci for 48 strains, establishing that VNTR typing could be used successfully for the genetic discrimination of *M. tuberculosis* strains (Figure 5). Later on, several publications

evaluated the polymorphism of additional loci (Mazars *et al.*, 2001; Le Flèche *et al.*, 2002), and proposed assays with increased discriminatory power and phylogenetic value. The advantage of MLVA compared to the previous typing methods is its relative ease of use, its reproducibility in different settings and with very different technological equipments (agarose gel, less expensive to set-up, but requiring more technical expertise; or capillary electrophoresis), its discriminatory power and informativity of its clustering. Nowadays, MLVA and spoligotyping are often used in combination as first-line assays, before running whole genome draft sequencing on the fraction of relevant strains. An MLVA assay comprising 24 loci is commercialized world-wide by a French company.

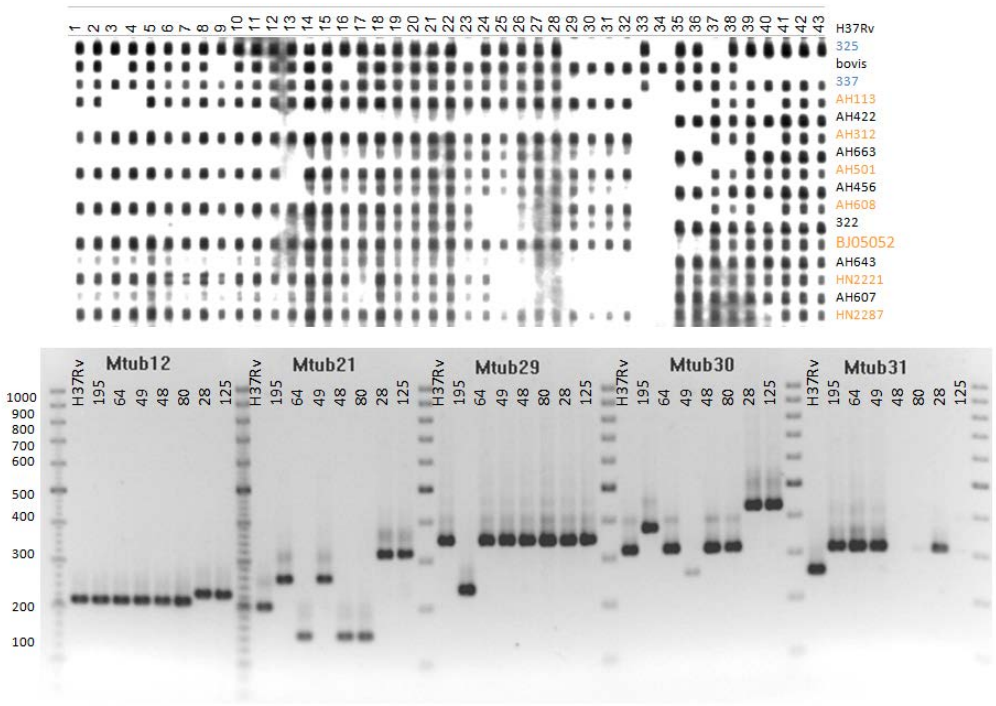


Figure 5: Examples of raw typing results.

Top, spoligotyping, hybridization patterns on a membrane; bottom, MLVA, results of the PCR amplification of 5 VNTRs. On the left, bands of the size marker are indicated in base-pairs (a mix of 100bp ladder and 20bp ladder).

Molecular genetics typing tools are aimed at identifying and understanding outbreaks. This goal was hard to meet with the first molecular typing tools as they did not have a sufficient resolution. The advent of high-throughput sequencing enabled to

access a level of resolution yet unknown, permitting to consider in more details the relationship between the strains from the same outbreak.

One of the first such studies was released in 2011 and was based on an epidemic from British Columbia (Canada). Thirty-two strains (29 from the same MLVA clone / outbreak and 3 older samples) were sequenced and put in relation with a social-network analysis of the interactions between the infected people during 2006. RFLP and MLVA identified the same pattern for all the strains, confirming the hypothesis of a clonal outbreak (Gardy *et al.*, 2011). This clone was also identified in previous cases that had occurred in the same area since 1995. The conclusion of the combination of the social analyses and the SNPs data led the authors to define two groups of strains that would share no particular relations, indicating the possible occurrence of two concomitant outbreaks (this clustering and interpretation was subsequently challenged). This study provided the community of TB researchers with a dataset of closely related strains from one epidemic event that could be used to calibrate the mutation rate, as the strains' sequencing reads were available and the duration of the epidemic was known.

1.5) The MTBC phylogeny.

The members of the *M. tuberculosis* complex are genetically very similar, showing on average 99.95% nucleotide similarity in whole genome comparisons. The split of the MTBC in different species reflects mostly host preference and some authors prefer to speak about specific ecotypes (*sensu* (Cohan)), as for instance in (Smith *et al.*, 2006). In this description the ecotypes differ mostly by their host adaptation, with for example *M. bovis* representing the clade infecting mainly cattle. The lineages inside *M. tuberculosis sensu stricto* could also be envisioned as ecotypes as they show a geographic distribution preference which might be related to the human genotypes they infect (Gagneux *et al.*, 2007).

Reconstructing the evolution of the *M. tuberculosis* complex may help understand how this very successful human pathogen emerged. During the last decade, several teams have performed phylogenetic studies. They used the molecular

tools developed for epidemiological purposes, as well as new ones, in order to propose different models to explain the evolutionary history of *M. tuberculosis*. These scenarii can be divergent on some aspects, but they illustrate the construction of a global consensus, and the steps needed to obtain the formulation of the current model for the emergence of the MTBC.

In 2002 Brosch and col. studied insertion/deletion events on the genome of 100 MTBC strains (Brosch *et al.*, 2002). The first conclusion of this analysis was that these events were unique, and could be used to characterize lineages in the complex. The deletions which took place in the progenitor strain giving rise to a specific lineage, were maintained in the lineage genomes due to the absence of recombination events in this clonal pathogen. The major deduction made from this analysis was that, contrary to what was considered as the most likely hypothesis at that time, *M. tuberculosis* could not have emerged from *M. bovis*. The *M. bovis* genome was smaller and presented more deletions. The different deletions identified on the collection of strains permitted to define subgroups. Two major deletions enabled to differentiate between *M. tuberculosis* and the other subspecies of the complex: the TbD1 region was deleted in all the *M. tuberculosis* strains except those of a so-called "ancestral" lineage, whereas the RD9 region was deleted only in *M. africanum* (human pathogen) and in the animal-adapted lineages, but not in the *M. tuberculosis* strains (Figure 6). One caveat to consider when interpreting the results of this study is that it used a limited number of sequenced genomes then available. The identified deletions pinpointed a number of key evolutionary events along a linear tree going from *M. tuberculosis* reference strain H37Rv to reference strain *M. bovis* BCG. One strain from "*M. canettii*", pathogen closely related to *M. tuberculosis* (van Soolingen *et al.*, 1997) (see dedicated paragraph below), was positioned as an outgroup in this study. This is because it is intact for both TbD1 and RD9 (like the "ancestral lineage"), and it happened to have a deletion overlapping with RD12 (Figure 6). In addition, it displayed numerous polymorphisms in some housekeeping genes.

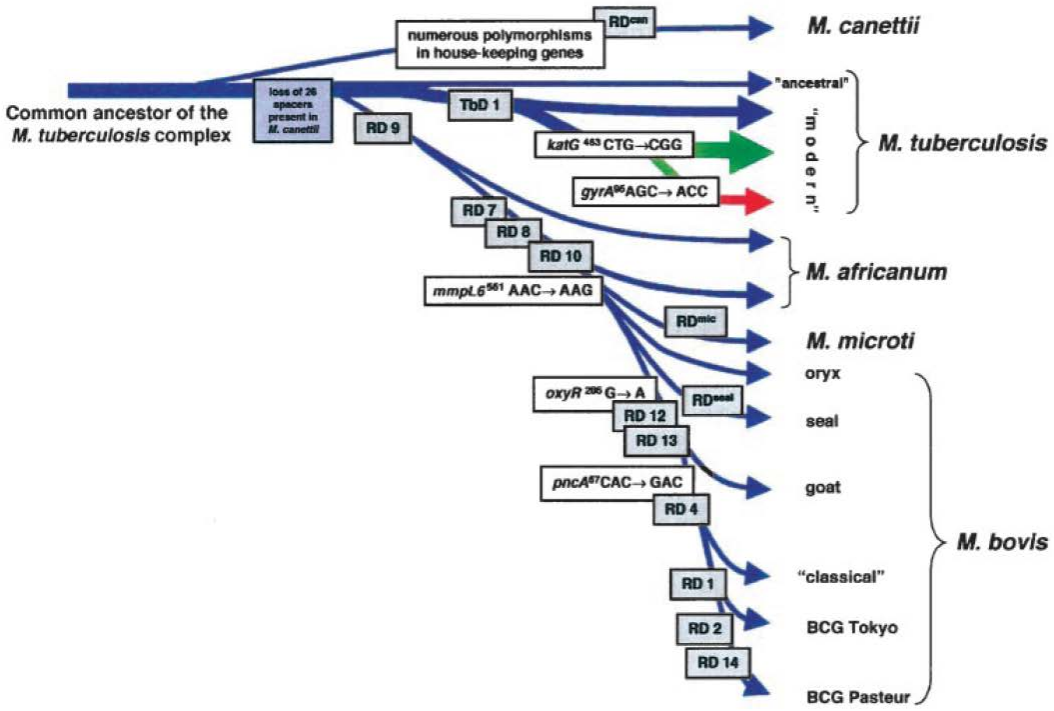
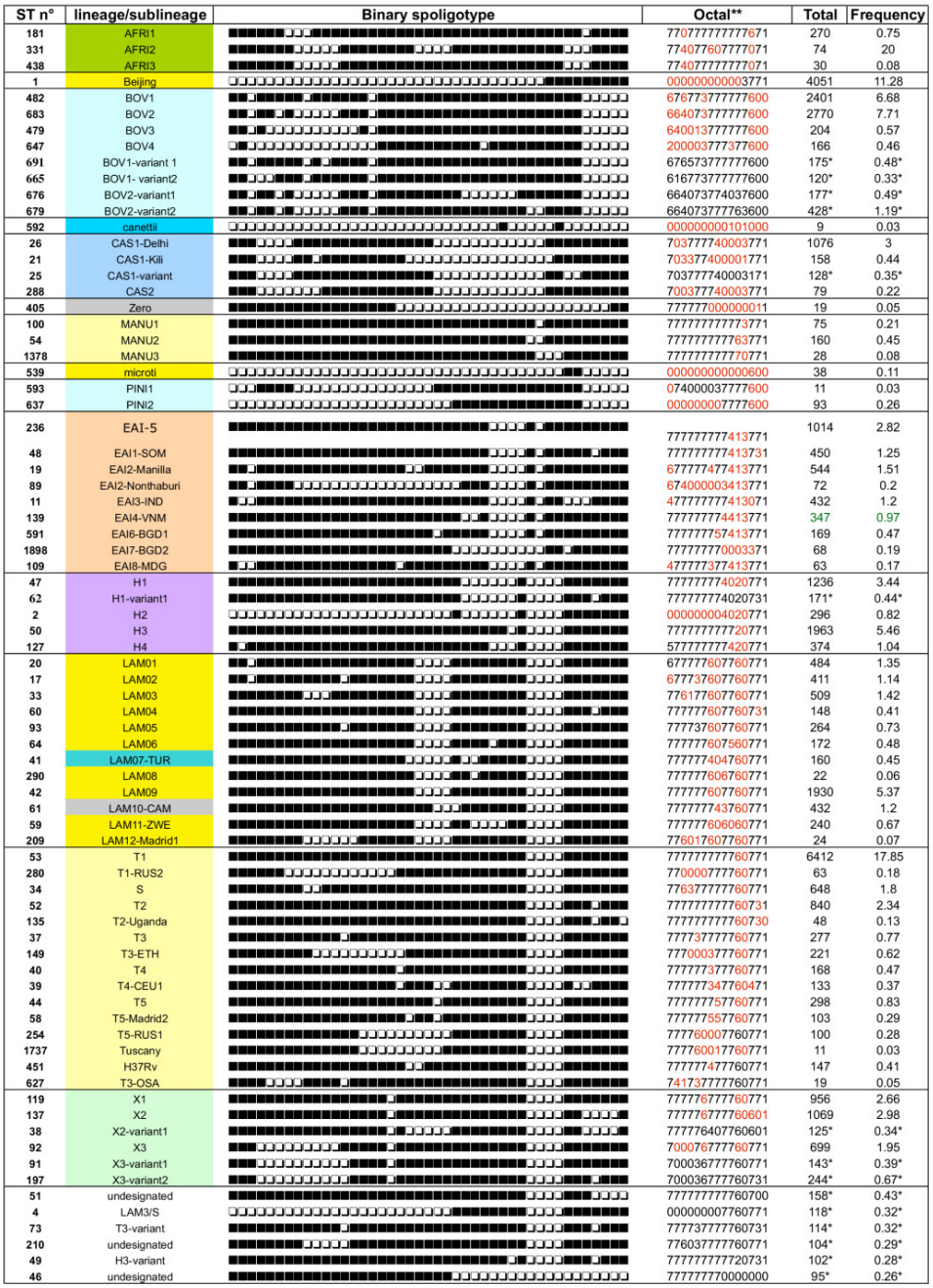


Figure 6: Phylogenetic tree based on deletions.

Reproduction of figure 2 from (Brosch *et al.*, 2002), copyright (2002) National Academy of Sciences, USA. This tree summarizes the model of evolution developed from the analysis of the deletions identified along the branch leading from *M. bovis* to *M. tuberculosis* H37Rv (gray boxes). The different animal lineages are detailed (different host preference).

The same year Filliol and col. made use of the data gathered from spoligotyping in order to describe the MTBC phylogeny (Filliol *et al.*, 2002). The authors used the spoligotypes gathered in the version of the spoligotypes database, then named SpolDb3, that has now evolved into SITVIT (cf. section 1.4). The first step of their analysis was to apply data-mining techniques on this dataset in order to obtain a clustering of the strains analyzed, leading to the definition of a number of sub-clades inside the complex. They also considered the links existing between these groups and the geographic distributions of the strains, and saw that several sub-families were preferentially found in certain regions of the globe. This aspect was detailed the following year (Filliol *et al.*, 2003), identifying 36 major clades and describing their phylogeographic repartition. The authors also used this classification to define nine major families that are now integrated into the six commonly considered lineages of the *M. tuberculosis* complex (namely *M. africanum* (lineages 5 &6), Beijing (lineage 2),

M. bovis (lineage 6), EAI (lineage 1), CAS (lineage 3), and T, Haarlem, X & LAM (lineage 4)). This study was updated in 2006 (Brudey *et al.*, 2006) (Figure 7) and 2012 (SITVIT Web, (Demay *et al.*, 2012)).



ST n°	lineage/sublineage	Octal**	Total	Frequency
181	AFRI1	77077777777671	270	0.75
331	AFRI2	774077607777071	74	20
438	AFRI3	77407777777071	30	0.08
1	Beijing	000000000003771	4051	11.28
482	BOV1	676773777777600	2401	6.68
683	BOV2	664073777777600	2770	7.71
479	BOV3	640013777777600	204	0.57
647	BOV4	200003773777600	166	0.46
691	BOV1-variant 1	676573777777600	175*	0.48*
665	BOV1- variant2	616773777777600	120*	0.33*
676	BOV2-variant1	664073774037600	177*	0.49*
679	BOV2-variant2	664073777763600	428*	1.19*
592	canettii	000000000101000	9	0.03
26	CAS1-Delhi	703777740003771	1076	3
21	CAS1-Kili	703377400001771	158	0.44
25	CAS1-variant	703777740003171	128*	0.35*
288	CAS2	700377740003771	79	0.22
405	Zero	777777000000011	19	0.05
100	MANU1	777777777773771	75	0.21
54	MANU2	777777777763771	160	0.45
1378	MANU3	777777777770771	28	0.08
539	microti	000000000000600	38	0.11
593	PINI1	074000037777600	11	0.03
637	PINI2	000000007777600	93	0.26
236	EAI-5	77777777413771	1014	2.82
48	EAI1-SOM	777777777413731	450	1.25
19	EAI2-Manilla	677777477413771	544	1.51
89	EAI2-Nonthaburi	674000003413771	72	0.2
11	EAI3-IND	477777777413071	432	1.2
139	EAI4-VNM	777777774413771	347	0.97
591	EAI6-BGD1	777777757413771	169	0.47
1898	EAI7-BGD2	777777770003771	68	0.19
109	EAI8-MDG	477777377413771	63	0.17
47	H1	77777777774020771	1236	3.44
62	H1-variant1	777777774020731	171*	0.44*
2	H2	000000004020771	296	0.82
50	H3	777777777720771	1963	5.46
127	H4	577777777420771	374	1.04
20	LAM01	677777607760771	484	1.35
17	LAM02	677737607760771	411	1.14
33	LAM03	776177607760771	509	1.42
60	LAM04	777777607760731	148	0.41
93	LAM05	777737607760771	264	0.73
64	LAM06	777777607560771	172	0.48
41	LAM07-TUR	777777404760771	160	0.45
290	LAM08	777777606760771	22	0.06
42	LAM09	777777607760771	1930	5.37
61	LAM10-CAM	777777743760771	432	1.2
59	LAM11-ZWE	777777606060771	240	0.67
209	LAM12-Madrid1	776017607760771	24	0.07
53	T1	77777777760771	6412	17.85
280	T1-RUS2	770000777760771	63	0.18
34	S	776377777760771	648	1.8
52	T2	777777777760731	840	2.34
135	T2-Uganda	777777777760730	48	0.13
37	T3	777737777760771	277	0.77
149	T3-ETH	777000377760771	221	0.62
40	T4	777777377760771	168	0.47
39	T4-CEU1	777777347760471	133	0.37
44	T5	777777757760771	298	0.83
58	T5-Madrid2	777775577760771	103	0.29
254	T5-RUS1	777760007760771	100	0.28
1737	Tuscany	777760017760771	11	0.03
451	H37Rv	777777477760771	147	0.41
627	T3-OSA	741737777760771	19	0.05
119	X1	77777677777760771	956	2.66
137	X2	777776777760601	1069	2.98
38	X2-variant1	777776407760601	125*	0.34*
92	X3	700076777760771	699	1.95
91	X3-variant1	700036777760771	143*	0.39*
197	X3-variant2	700036777760731	244*	0.67*
51	undesignated	777777777760700	158*	0.43*
4	LAM3/S	000000007760771	118*	0.32*
73	T3-variant	777737777760731	114*	0.32*
210	undesignated	776037777760771	104*	0.29*
49	H3-variant	777777777720731	102*	0.28*
46	undesignated	777777770000000	95*	0.26*

Figure 7: Spoligotypes of the different lineages and subgroups of the MTBC.

Reproduction of figure 1 from (Brudey *et al.*, 2006). Spoligotypes are grouped by clades, showing the binary pattern (black and white squares) and the corresponding octal code. The numbers appearing in the right columns correspond to the content of SpolDb4 at the time of the publication (2006).

In 2006 Filliol *et al.* used 212 Single Nucleotide Polymorphisms (SNPs) to analyze 323 strains (Filliol *et al.*, 2006). This SNPs set had been determined by comparing the then publicly available whole-genome of four *M. tuberculosis* strains; the authors selected the SNPs that were characteristic of one of the genomes (there is no SNP shared by two genomes out of the four considered for the determination of the dataset). These 212 SNPs were then assessed in the 323 studied strains and the results were used to construct a phylogeny, leading to the definition of six major genetic groups. The analysis of these groups showed that they were largely correlated with the spoligotypes of the considered strains, confirming the definition of the major lineages deduced from spoligotyping. Besides, the authors used the tree generated with the SNPs to infer some evolutionary hypotheses regarding the different groups of the complex. They identified the EAI (East Africa and India) lineage as the most ancestral, and therefore postulated that the MTBC could have emerged in India, and then have spread to the rest of the world through East Asia.

Two years later Wirth *et al.* used VNTR typing (24 loci panel) on a collection of 355 *M. tuberculosis* strains to characterize the phylogenetic lineages inside the complex (Wirth *et al.*, 2008). The authors defined two main clades regrouping all the *M. tuberculosis* strains, except the members of the EAI lineage (a.k.a. lineage 1 in the current nomenclature) on one hand, and all the *M. africanum* and the animal-adapted strains on the other hand (lineages 5 & 6), according to the deep branching position of these clades in the phylogeny. The exact position of the EAI lineage was somewhat unclear in this study. The authors then used short term mutation rate estimates of VNTR loci, and Bayesian statistical analysis to try to put dates on this phylogeny. They obtained an estimate of around 40,000 years BP for the most recent common ancestor (MRCA) of the MTBC, with estimates for the main nodes inside the tree. They proposed a model that postulated an emergence in the Horn of Africa with a subsequent spread to the Fertile Crescent, and then a "return" of one clade in Africa while the other major clades spread in the rest of the world. Finally they used the MLVA typing data to link the current observed diversity with an increased spread of tuberculosis world-wide in conjunction with the Industrial Revolution.

The same year Hershberg *et al.* compared 89 gene sequences (65,829bp) from 108 diverse strains (as estimated by spoligotyping) (Hershberg *et al.*, 2008). This search of SNPs by scanning 1.5% of the *M. tuberculosis* genome, instead of larger polymorphisms identified by comparing two genomes, enabled the authors to access a greater resolution and, most importantly, an unbiased view for their analysis of the branching points. The authors proposed an evolutionary scenario, with some attempt at dating the major changes in the phylogenetic history of the complex. They concurred in the sense of an Out-of-Africa emergence model. Using circumstantial evidence, as *M. tuberculosis* seems to present some genetic adaptations to a low-density setting that could be consistent with early human hunter-gatherers populations, they dated it back to -50,000 years. They concluded that the contemporary distribution of lineages seemed to coincide with the early movements of human populations, so that the diversification could have happened approximately during the same time frame. They proposed that the initial steps in the dispersion of the MTBC would have taken place by land, with the first human migrations, and that, in the last few centuries, the different lineages would have mixed due to the increase in the movements of human populations, as for instance owing to the development of travel by sea since the XVth century.

With the advent of high-throughput sequencing technologies (see following technical section), the use of genome wide SNPs detection for phylogenetic investigations became less expensive than the partial sequencing of a subset of genes. This approach was applied in 2010, for 22 strains representing the lineages defined at this time (Comas *et al.*, 2010). A total of 9,000 SNPs were retained to construct a detailed phylogeny describing the relations between the different strains. This study attempted to determine if essential genes were less susceptible to mutation than non-essential ones, which is a sign of purifying selection, but there was no clear evidence for this. The results confirmed that, in general, the dN/dS ratio was higher than in other bacteria of the same order. This analysis was focused on the study of genes recognized by the human T cells during infection. It appeared that, contrary to what could have been expected in order for the bacteria to escape the human defense

mechanisms, these loci were well-conserved, meaning that the recognition must have no adverse effects for the pathogen inside its host.

1.6) Horizontal gene transfer and mutation rate.

Available molecular data emphasize that the genetic structure of the complex is clonal. The different markers used to cluster the MTBC members did not show detectable signs of horizontal gene transfer (HGT) between or within lineages. Yet in other pathogenic bacteria, as well as in the closely related "*M. canettii*" strains (cf. sections 1.7-1.8), HGT has been identified, and sometimes plays a key role in pathogenicity (as for instance in the case of pathogenicity islands). In 2007, the genomes of three MTBC strains were analyzed in order to identify traces of putative HGT events (Becq *et al.*, 2007). Out of 4.4Mb of genomic sequences, about 200kb could be interpreted as resulting from HGT. The study of the same region in the "*M. canettii*" strains indicated that these events were already present in the progenitor species of the MTBC and were consequently quite ancient. The putative origin of this genetic material included soil bacteria and potential animal pathogens.

Another element of interest for population genetics analyses is the mutation rate that characterizes a particular organism in specific environmental conditions. If this mutation rate could be perfectly determined, the evolutionary time between two strains from their MRCA could be precisely determined from the number of mutations between them. However selection impacts the "natural" mutation rate by providing constraints on the way mutations will be conserved or purged from a population, and those constraints can vary widely depending on the environmental conditions. For *M. tuberculosis*, it was shown in a macaque model that the mutation rate per day was the same in any of the disease states (i.e. active disease or latent infection), and also during culture in liquid medium (Ford *et al.*, 2011). A high proportion of mutations could be linked with oxidative stress caused by the presence of the bacteria in the macrophage's phagosome.

1.7) About "M. canettii": discovery and pathogenicity.

"*Mycobacterium canettii*" is the name given to a mycobacterium isolated from a clinical sample of a supposedly French farmer presenting tuberculosis-like symptoms in 1969 by Georges Canetti, from the Pasteur Institute. The strain apparently belonged to the MTBC but possessed phenotypic characteristics that set it aside classical members of this complex. The most striking one was that when cultured on solid medium it formed smooth colonies, whereas "normal" MTBC members formed typical rough colonies (Figure 8). This group of strains was therefore named "*Mycobacterium tuberculosis* strain Canetti". However the exact nature and significance of this newly discovered strain remained elusive, as it was very rare. Its true phylogenetic relationship to the MTBC was then undetermined.

Nearly 20 years later, in 1987, the lipidic composition of the cell wall of the four strains known at that time (CIPT 1400100-59, -60,-61 & -62) was characterized by a team of French researchers (Daffé *et al.*, 1987; Lemassu *et al.*, 1992). One major result of this study was that the "Canetti strain" had lipids that were close to some present in *M. bovis* and *Mycobacterium kansasii*, but not in *M. tuberculosis*. This, together with the phenotypic differences noted when they were first isolated, were indications that these strains were more peculiar than just *M. tuberculosis* strains forming strange colonies. The geographic origin of these four historical strains is not documented.

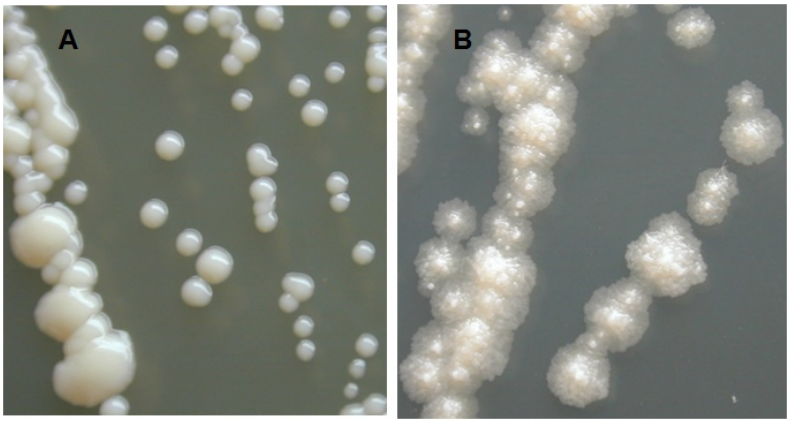


Figure 8 : Morphology of tuberculosis-causing colonies.

Reproduction from (Koeck *et al.*, 2011). A, smooth "*M. canettii*" colonies, B, typical rough MTBC member colonies.

The next study on "smooth *M. tuberculosis*" (in reference to the phenotype that enabled their identification) was published ten years later after a new strain was isolated by researchers from the Netherlands (van Soolingen *et al.*, 1997). This study was the first to establish more precisely some phylogenetic characteristics of the smooth strains. The new strain came from a two-year-old Somalian child with lymphadenitis. It had the same lipid characteristics as those described by Daffé *et al.*. Several genetic markers differentiated it from members of the *M. tuberculosis* complex, as for instance IS1081 copy number (another frequently used insertion sequence) or RFLP profile. Its spoligotype was also fairly different from what had been observed in the complex. When compared with one of the Canetti strains it appeared that these two strains had several characteristics in common. The authors proposed that they be treated as a new subspecies in the MTBC, which they called "*Mycobacterium canetti*". These strains had a shorter generation time than the reference *M. tuberculosis* strain H37Rv, and they showed the same level of virulence in guinea pigs. One major achievement of this study was the identification of clustered SNPs in the *recA* genes of the smooth strains, which proved that, contrary to the clonal MTBC, smooth strains were submitted to recombination with an unknown donor.

Figure 9: Map of the Horn of Africa.

The following year, another publication described a new smooth strain, from a Swiss patient suffering from mesenteric tuberculosis, who had been living in Uganda and Kenya before coming back to Switzerland (Pfyffer *et al.*, 1998). The patient developed the first signs of the infection while in Kenya, and was HIV-positive. As the strain from this patient formed smooth and glossy colonies, it was further investigated to know if it belonged to the newly defined "*M. canettii*" subspecies (the authors modified "*M.*

canetti" in "*M. canettii*" in agreement with Latin and taxonomic rules)). The IS6110 RFLP patterns was identical to those of the Somali strain described by van Soolingen *et al.* (van Soolingen *et al.*, 1997). This proximity between the two strains was also confirmed by spoligotyping. The discoveries of these two strains began to hint at a key element that would thereafter be considered like one of the main characteristic of "*M. canettii*", namely its probable geographical restriction to the East of Africa, in a region somewhere near Kenya and Somalia.

This localization was strengthened by the results of French researchers some time later. The French Army Health Services identified two new cases of infections caused by "*M. canettii*" in two soldiers, from the French Foreign Legion based in Djibouti (Republic of Djibouti, cf. map Figure 9) where the French Army has a permanent establishment. Typing by RFLP and spoligotyping identified these strains as "*M. canettii*", confirming what the smooth aspect of the colonies already seemed to indicate on a morphological level (Miltgen *et al.*, 2002). The two patients suffered from pulmonary tuberculosis; this study was the first to report "*M. canettii*" infections in "classical" pulmonary tuberculosis. The authors concluded their study by stating that infection by the smooth tubercle bacillus seemed to be clearly geographically correlated with a stay in the region of the Horn of Africa. In 2004, close to forty smooth strains had been reported (Fabre *et al.*, 2004).

It is not before 2011 that extensive clinical studies of smooth strains were published. In a collaboration of our laboratory with the French Army Health Services, thirty additional "*M. canettii*" strains were isolated during a 3-years period, accounting for nearly 10% of all TB infections in the involved hospitals from Djibouti (Fabre *et al.*, 2010; Koeck *et al.*, 2011). The microbiological study confirmed that the smooth strains grow twice as fast as the classical MTBC members in liquid medium, even if the difference did not appear to be as significant on solid medium. From a clinical point of view it appeared that the patients infected by "*M. canettii*" were significantly younger than those infected by rough strains. French expatriates are more prone to be infected by smooth strains as compared to Djiboutians. The signs of the infection were not different from an infection caused by a member of the *M. tuberculosis* complex, and

there were cases of pulmonary and extra-pulmonary tuberculosis (mainly lymph nodes infections in the second case) (Table 1). The last conclusion of this article was that no inter-humans transmissibility could be identified. Some smooth strains are able to infect humans, but they would not be contagious once a patient is infected, which is a major difference compared to the *M. tuberculosis* complex members. This was a strong argument for the existence of an environmental reservoir leading to the contamination of patients.

Strain ID	Patient nationality [a]	Age	Gender	Year of isolation	TB type [b]	MLVA group [c]
Percy3a	DJ	8	f	1998	Pulm	A
Percy6	FR	35	f	1998	Pulm	A
Percy8	FR	4	f	1998	LN	A
Percy21a	FR	36	m	1998	Pulm	A
Percy21b	DJ	UNe	f	2000	UN	A
Percy26	DJ	18	f	1999	LN	A
Percy29a	ETH	34	f	1999	Pulm	A
Percy74	DJ	UN	m	2001	Pulm	A
Percy99c	DJ	22	m	1999	Pulm	A
Percy150	DJ	UN	m	2000	Pulm	A
Percy156	DJ	27	m	2001	LN	A
Percy189b	DJ	UN	m	2000	Pulm	A
Percy199b	DJ	UN	f	2000	PerL	A
Percy205	DJ	63	m	2000	Pulm	A
Percy206	DJ	UN	m	2000	Pulm	A
Percy212	DJ	32	m	1998	LN	A
Percy245b	DJ	UN	f	2000	LN	A
Percy246	DJ	UN	m	2001	LN	A
Percy257	DJ	40	f	1998	LN	A
Percy32	DJ	13	m	1999	LN	
Percy79	DJ	UN	m	2001	Pulm	
Percy94	DJ	42	m	1999	Pulm	B
Percy213	DJ	20	m	1998	Pulm	B
Percy214	DJ	27	f	1998	Pulm	B
Percy258	DJ	40	m	1998	Pulm	B
Percy25	DJ	7	m	2000	LN	
Percy65	DJ	4	f	1999	LN	
Percy89	ETH	UN	f	2000	LN	
Percy99b	DJ	UN	m	2000	Pulm	
Percy157	DJ	UN	m	1999	Pulm	

Table 1: Details about the "*M. canettii*" strains of the collection (extract).

Modification of table 4 from (Koeck *et al.*, 2011). a - Patient nationality: DJ, Djibouti, FR, France, ETH, Ethiopia, ERI, Eritrea, SOM, Somalia. b - Pulm, pulmonary, LN, lymph node, PerL, peritoneal liquid. c - As defined in (Fabre *et al.*, 2004).

1.8) Phylogenetic analysis and relationship with the MTBC.

As smooth strains were described and became more numerous, their precise phylogenetic relationship with the *M. tuberculosis* complex became of interest for researchers working on the evolutionary history of the tubercle bacillus. These smooth strains that could cause tuberculosis, but had different phenotypic characteristics, could provide useful insight on the origins of the MTBC.

In 2000 van Embden *et al.* investigated the DR locus composition in a "*M. canettii*" strain, to complete their study of this locus in the MTBC (van Embden *et al.*, 2000). The strain was the one described by van Soolingen *et al.* from the Somali child. Twenty-six spacers were identified by sequencing as being specific of the smooth strain. None of the spacers found in the MTBC members were present in the "*M. canettii*" DR locus, although an artifactual signal on two spacers could be observed by spoligotyping. These observations were confirmed with the other smooth strains available at that time. With such a small number of smooth strains, the evaluation of the diversity was difficult, but it showed that the commonly used spoligotyping assay could not discriminate between smooth strains since the set of spacers present in "*M. canettii*" was not included in the assay.

In their study of deletions to characterize the phylogeny of the MTBC (cf. section 1.5), Brosch *et al.* included 5 smooth strains from the Pasteur Institute tuberculosis Collection (Brosch *et al.*, 2002). The smooth strains bore none of the deletions identified in the *M. tuberculosis* complex; particularly they were intact for TbD1 and RD9, like the EAI lineage. The presence of an apparently new deletion ($RD12^{CAN}$), the constitution of the DR region, with spacers absent from the rest of the MTBC, and the cluster of SNPs observed in the RecA gene prompted these authors to place "*M. canettii*" as a likely outgroup regarding the complex, even if it was still unclear if it had to be considered as a subspecies of *M. tuberculosis* or not. However as only five smooth strains were included in this work, no conclusion was established about the relationships between them.

In 2004 and 2005 two studies characterized the extended collection of smooth strains, *alias* "*M. canettii*", assembled by the French Military Health Services (called the SSA for "Service de Santé des Armées" collection) (Fabre *et al.*, 2004; Gutierrez *et al.*, 2005). Both studies made use of MLVA typing complemented by CRISPR locus analysis and partial sequencing of housekeeping genes in order to establish the relationship between these strains, and their link with the MTBC. They established that the *M. tuberculosis* complex was genetically more homogeneous than these related strains, and proposed that the MTBC emerged from a smooth-strain ancestor. Based on partial sequencing of six house-keeping genes, the 2005 study went on to propose that the progenitor species may be 50 to 100 times older than the MTBC. The smooth strains could be resolved into several distinct clusters, with one cluster comprising the majority of the strains isolated from patients in Djibouti during the gathering time, named cluster A in the 2004 study, and D in the 2005 article. The study of the DR locus revealed that, contrary to the MTBC members, some "*M. canettii*" strains seemed to be devoid of it, which once again hinted at more diversity than what had been observed inside the complex.

Gutierrez *et al.* confirmed by more extensive analyses what had been identified by van Soolingen and col. in the *recA* gene, namely the fact that the smooth strains undergo recombination with an unknown source (Gutierrez *et al.*, 2005). For this team, the heterogeneity observed when studying the smooth strains implied that they could not be considered as all belonging to "*M. canettii*", as defined by van Soolingen *et al.* in 1997. In particular the absence of a DR locus in some strains, as well as the presence in some strains of an intact $RD12^{CAN}$ region that was previously considered as characteristic of the "*M. canettii*" strains, led them to propose to distinguish the smooth strains in different subgroups. All smooth strains would be called "*Mycobacterium prototuberculosis*" and "*M. canettii*" would represent a subspecies of this newly named species. "*M. prototuberculosis*" would represent the progenitor species from which the *M. tuberculosis* complex, as well as "*M. canettii*", would have emerged at some point in the past. In their acceptation of genomic diversity, MTBC members and smooth strains should belong to a single bacterial species. The authors

concluded that the smooth tubercle bacilli may have been causing tuberculosis in humans for a much longer time than previously suspected. However, N. Smith underlined that the impact of recombination had not been correctly taken into account in the dating, and that available data did not yet prove that "*M. canettii*" was an outgroup of the MTBC complex (Smith, 2006).

Between 2004 and 2010, 30 new smooth strains were isolated in Djibouti. The complete collection of strains from this area was submitted to a detailed molecular analysis (Fabre *et al.*, 2010). This new analysis confirmed previous findings on the genetic diversity inside "*M. canettii*", as well as its apparent distance to the MTBC. It also refined the subgroups previously described in 2004 (Fabre *et al.*, 2004), by identifying two major genetically homogenous clones that seemed to be emerging in the Djibouti area (Figure 10). In particular the spacers specific to the "*M. canettii*" strains were determined. In their conclusion the authors indicate that the further understanding of the genetic diversity of the smooth strains would be achieved by the access to the whole genome sequence of some strains. This was the PhD project I was assigned when I arrived in the laboratory.

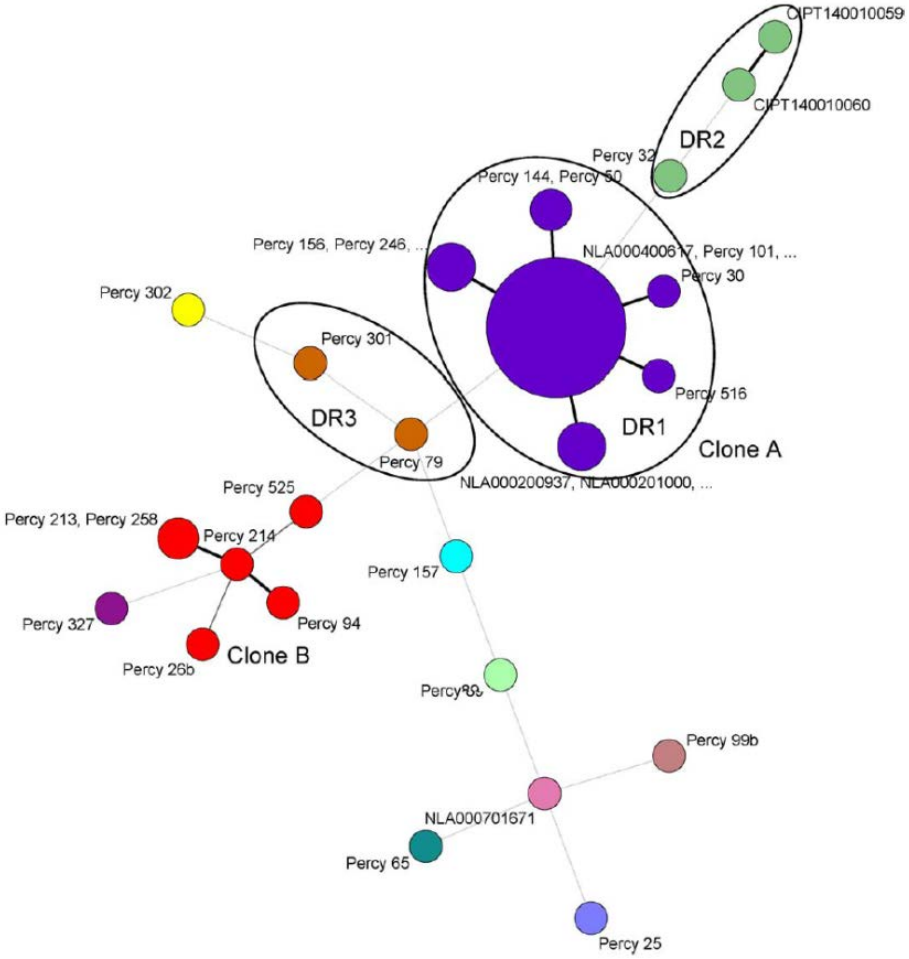


Figure 10: "*Mycobacterium canettii*" diversity.

Minimum Spanning Tree based on MLVA data, reproduced from (Fabre *et al.*, 2010). 59 strains were typed using 24 VNTR loci; the nodes' size is proportional to the number of strains in that node. The color code is used to indicate clusters of closely related genotypes. Three groups of strains possessing a particular CRISPR alleles are circled (DR1, DR2, DR3).

1.9) Environmental Mycobacteria.

To understand *M. tuberculosis* and its emergence it may be useful to consider its closely related neighbors in the genus *Mycobacterium*. Mycobacteria are often differentiated into two groups, namely the fast-growers and the slow-growers (Figure 11). Empirically this division can be determined with the appearance of visible colonies on culture medium in less than seven days for the fast growers, and more for the slow growers. The most pathogenic belong to the slow growers, whereas the fast-growers are usually environmental bacteria. This difference in growth rate could partly be

correlated to a difference in the permeability of the cell wall, but this question is still opened to debate (Hett *et al.*, 2008).

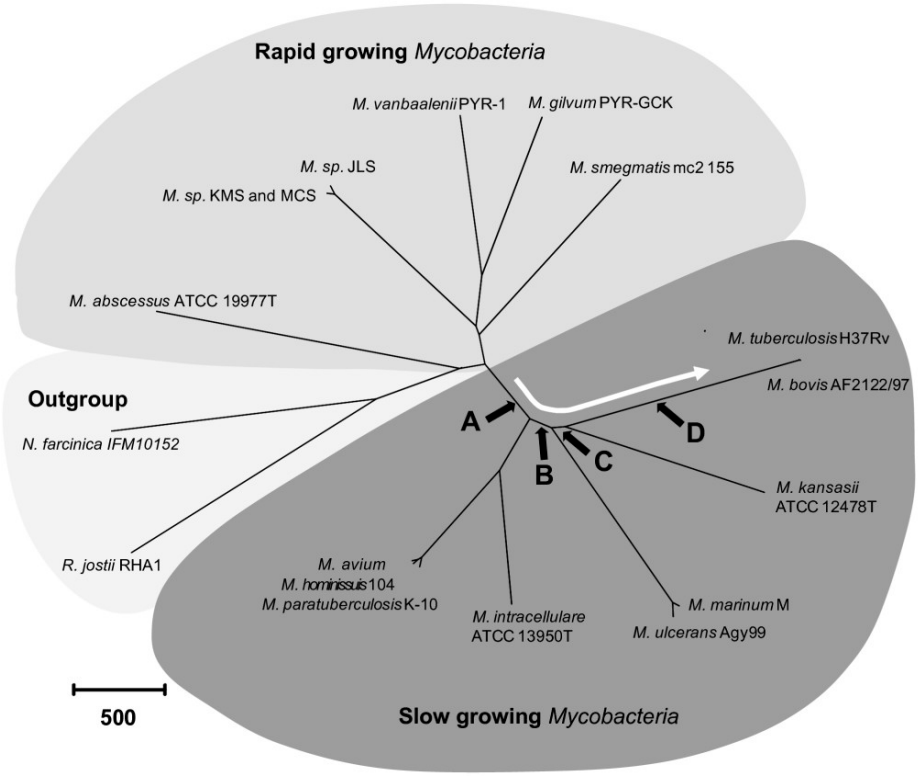


Figure 11 : Global tree of Mycobacteria

Radial tree reproduced from (Veyrier *et al.*, 2009), showing the different mycobacterial lineages. This tree is MLSA-based, using the genomic data available for the different lineages shown in the tree.

The two environmental Mycobacteria closest to the *M. tuberculosis* complex outside the smooth strains are *Mycobacterium marinum* and *M. ulcerans*. The latter, as the tubercle bacillus, is pathogenic for humans, causing Buruli ulcer, the third cause of mycobacterial infections after tuberculosis and leprosy. *M. ulcerans* has been discovered in 1948, but the exact reservoir of this bacterium is still mainly unknown. It is supposed to be linked with water environments, and possibly the presence of aquatic insects such as water bugs (Portaels *et al.*, 1999; Marsollier *et al.*, 2002). The infection could occur via the bite of a contaminated insect, or through a contact with contaminated insect residues.

M. marinum is also living in water and was first isolated in 1926 from a fish. From genetic comparison of the 16S RNA it appears that *M. ulcerans* and *M. marinum* are closely related, with 99.6% similarity. Their genome size is quite similar, even if that of *M. ulcerans* is slightly shorter (5.8 and 6 Mb, respectively). In a multiple loci sequence analysis (MLSA) study, Stinear *et al.* suggested that *M. ulcerans* could have emerged from a *M. marinum* ancestor (Stinear *et al.*, 2000). In this article the authors proposed that *M. ulcerans* could be considered as a peculiar ecotype of *M. marinum*, a fish pathogen that can sporadically infect humans in contact with diseased fish (such as aquariophils). When it infects humans, it affects mainly the extremities of the body (because of the origin of the infection which involves mostly the skin) (Stamm *et al.*, 2004). The disease produces granulomas, reminiscent of an infection by *M. tuberculosis*. As *M. marinum* is one of the environmental mycobacteria closest to the MTBC, it has been an object of study in order to find significant genetic characteristics that could be used to study the members of the *M. tuberculosis* complex. Because its genome is 2Mb bigger than that of *M. tuberculosis,* it is thought that the MTBC could have emerged from an environmental bacterium that had an ecology close to that of contemporary *M. marinum* strains. Its genome would have then evolved by reduction, losing genetic elements that were mandatory for survival in the environment, and acquiring genes specific of the life cycle of an obligate intracellular pathogen. However, the genetic composition of the genome of these two species is quite different, as they have in average an amino-acid identity of 85% (Stinear *et al.*, 2008).

Another mycobacterium worth mentioning in this section is *Mycobacterium kansasii*, which causes a tuberculosis-like disease in humans and is also closely related to the MTBC. It causes a similar pulmonary disease, and has been isolated from tap water (Engel *et al.*, 1980). This is another example of a water-based mycobacterium that causes a pulmonary infection similar to tuberculosis, reminiscent of the relationship between *M. tuberculosis* and *"M. canettii"*.

For the purpose of understanding the emergence of the *M. tuberculosis* complex the interactions between mycobacteria and environmental amoeba have been studied by Salah *et al.* (Salah *et al.*, 2009). Mycobacteria are shown to be

amoeba-resistant organisms, i.e. they can survive and multiply inside the amoeba. This capacity may explain how the MTBC developed into an intracellular pathogen, capable of invading human macrophages. The biological mechanisms implied in the survival into amoeba are the same as those needed for survival in the macrophages, particularly the ones that impair the fusion of phagosomes with lysosomes. Furthermore, the natural environment of amoeba is water, where the mycobacteria closest to the complex have been found and seem to live, which strengthen the potential link between the bacteria and these unicellular eukaryotes.

1.10) Historical records for tuberculosis.

Before the nineteenth century and the successful efforts to discover the natural causes of tuberculosis, not much was known about this disease. However owing to historical records describing the symptoms associated with tuberculosis, it is possible to try to infer the history of tuberculosis from the traces it has left on human history. Moreover, since a few decades, the field of paleopathology has developed tremendously, permitting to try and reconstruct the evolutionary history of a number of pathogens from the remains of their victims. The most recent development in that field is probably the advent of ancient DNA (aDNA) study, enabled in part by the high-throughput sequencing technologies, even if the field is still undergoing major evolutions.

In the nineteenth century, testimonies are plentiful, and it is quite easy to get a panorama of the extant of the tuberculosis epidemic at this time. Tuberculosis was a scourge in the poorest levels of society, as promiscuity and poor health habits lead to an easy transmission of the pathogen, and a high lethality. As recalled in 1.1, tuberculosis was also wide-spread in the artistic community, as can be seen in many artistic works from that period such as books or paintings, and was considered as nearly always fatal.

As we try to go back in time, the information becomes scarcer. It seems that tuberculosis was already quite prominent during the Middle Ages, but the reports are not as detailed as for the later time periods. Moreover, during this era, several other

pathogens like *Yersinia pestis* were already claiming a lot of lives, somewhat eclipsing those caused by tuberculosis. Nonetheless, it seems that the incidence of TB got up at the end of the Middle-Ages and during the Renaissance, possibly because of the growth of cities and the augmentation of population densities, linked with the Industrial Revolution (Chalke, 1959).

Several authors from the Antiquity mention a disease probably corresponding to tuberculosis, as for instance Greek writers like Herodotus or Hippocrates. TB was then called phthisis, which meant consumption. There seems to be mention of the same disease in ancient Chinese medical texts from the fifth century BCE (Elvin *et al.*, 1998), and also in India in 1,500 BCE (Zysk, 1998). In Ancient Egypt there are written traces of the existence of likely tuberculosis infection in the Pharaonic civilization. But all these elements are only second-hand proofs. What about "real" evidence of the presence of the tubercle bacillus?

Many researchers have tried to infer the presence of infection by MTBC members from the analysis of skeletal remains. Some physical changes supposedly associated with tuberculosis were analyzed, in order to study the prevalence of the disease in ancient times. However, until recently, physical observation was the only mean available to do this, even if it does not provide a formal proof. In the last decades the advent of molecular techniques began to open the field to more precise studies. It is now possible to amplify DNA from ancient human remains, in order to get access to the sequence of a bacterial pathogen having infected a particular individual. This has already been done successfully for several other pathogens (for instance for *Y. pestis* (Morelli *et al.*, 2010; Bos *et al.*, 2011; Wagner *et al.*, 2014)), as well as for MTBC members.

One of the first such studies for the MTBC was performed by Zink and col. in 2003 (Zink *et al.*, 2003). The authors studied DNA from mummies that had been buried in the necropolis of Thebes in Upper Egypt between 2,050 and 500 BC. They were able to retrieve aDNA from the mummified remains, and used it to perform several molecular typing analyses. They studied the presence of insertion sequence IS6110, as well as the spoligotyping profile. These methods enabled them to confirm the

presence of *M. tuberculosis* DNA and most importantly to attribute the strains to lineages defined in the literature. One of the main points of this publication is that they identified the absence of spacer 39 and the presence of spacer 8 in the most ancient samples (in a tomb that had been used until 1,600 BC), but not in more recent sepultures. These data suggest a switch between *M. africanum* and *M. tuberculosis* during the second millennium BC, as the ancient samples' spoligotype is characteristic of some *M. africanum* strains (cf. *M. africanum* spoligotypes, Figure 7).

Another study using aDNA was performed on remains from a Neolithic settlement from the Eastern Mediterranean and dated from 9,000 years before present (7,000 BC) (Hershkovitz *et al.*, 2008) (Figure 12). In this analysis, *M. tuberculosis* DNA was detected and characterized, either by study of insertion sequences (IS6110, IS1081) or by PCR amplification of specific locations on the *M. tuberculosis* genome. The resulting information enabled the authors to claim that the pathogen infecting the skeletons was a *M. tuberculosis* strain from a TbD1-deleted lineage. They also directly analyzed the samples in order to detect the presence of mycolic acids, which gave positive results, thereby confirming the presence of a mycobacterium in the studied samples. However due to the degraded quality of aDNA samples, the reproducibility of some analyses seems to be difficult, which prevents to draw any decisive conclusion from those results.

Tuberculosis has also been identified in skeletal remains from Germany dating back to 5,400 years BC (Nicklisch *et al.*, 2012). The identification is based both on analysis of osseous lesions and on aDNA amplification. IS6110 was used to assess the presence of MTBC DNA. To further improve this description, the TbD1 and RD9 regions of deletion were amplified, resulting in the discovery that the TbD1 region seemed to be intact whereas RD9 may have been already deleted as in the *M. africanum* lineage. However a more precise characterization, as for instance by spoligotyping, was not possible due to the highly degraded state of the genetic material. Another interesting point emphasized by the authors was the difference in pathological signs observed between the three sites included in their study. Indeed, for one of the sites, the skeletal remains showing a positive signal for aDNA identification, showed less osseous

lesions, except for the ribs, but at the same time presented biochemical characteristics of a default in protein uptake, a probable sign of malnutrition. This led the team to propose that, in this particular settlement, TB could have been a wide-spread chronic disease affecting many individuals, due to poor environmental conditions, mainly a lack of meat in this early farming community.

The oldest known trace of tuberculosis caused by a *M. tuberculosis* like bacterium came from the Americas, with the skeleton of a 18,000 years old bison. This skeleton presented physical alteration compatible with those known for bovine TB infection, and was therefore submitted to an extensive analysis in order to characterize the pathogen responsible for these lesions (Rothschild *et al.*, 2001). Amplification of aDNA was possible, and the determined spoligotype octal code was 7F-6E-7E-7F-F8-7D (missing spacers 10, 14, 21, 34, 35, 36 & 42). This profile does not correspond to any known spoligotype family (Figure 7). The authors suggest that this might be expected if this lineage is older than the MTBC's MRCA.

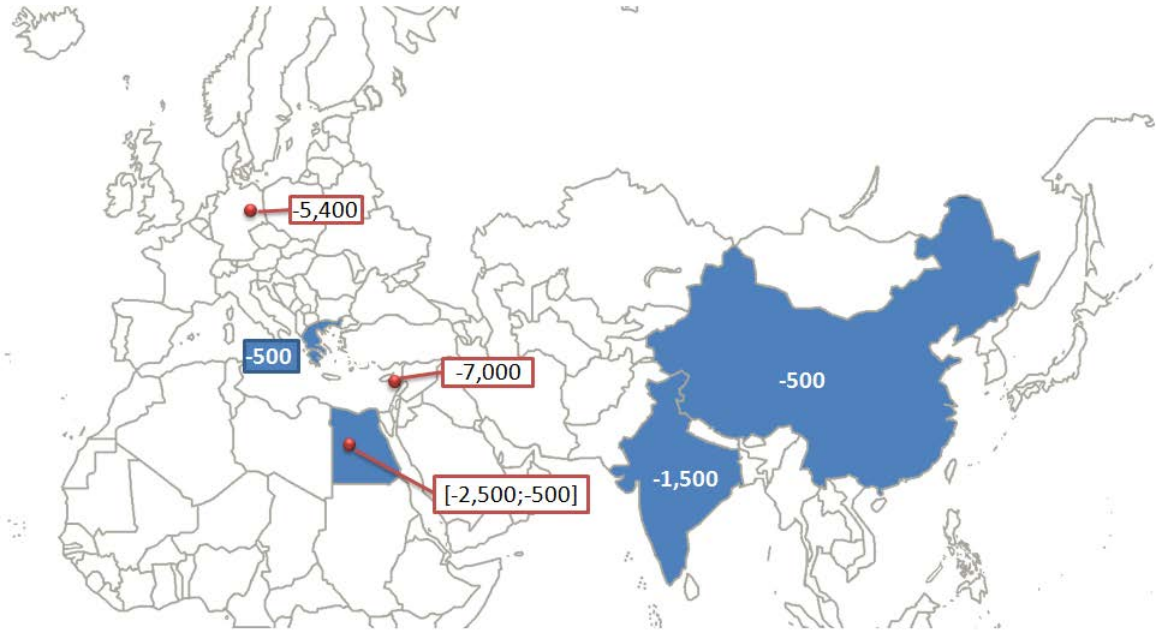


Figure 12: Map of the evidence for historical occurrences of tuberculosis during human history.

The blue areas correspond to literary sources, the red dots to ancient DNA studies where MTBC DNA has been identified.

1.11) History of the Horn of Africa.

As the currently most accepted model of emergence of the *M. tuberculosis* complex is an East-African origin (Fabre *et al.*, 2004), it seems useful to have a look at the prehistoric past of the East African region regarding human populations: co-evolution between the pathogen and its host would imply a high degree of correlation between Human history and the spread and diversification of the tubercle bacillus.

Anatomically modern humans (*Homo sapiens*) would have appeared in the East of Africa some 200,000 years ago. Ethiopia seems to be one major location for the emergence of *Homo sapiens* ((Fleagle *et al.*, 2008), site of Omo Kibish). Little is known about the earliest times of human occupation in East Africa (as well as in the whole world in general). Around 200,000 years ago, the global human population is estimated to be of about 20,000 individuals, and the average life expectancy is 15 years. Much closer to our era, 20,000 years ago the global human population is estimated at 3 million people, with an average life of 30 years. This means that until 20,000 years ago, approximately 7 million humans had lived (and died) on Earth since the emergence of modern humans. Furthermore it seems that due to climatic changes occurring around 70,000 years ago, the total population (at this time still mainly in Africa) could have dropped to around 10,000 individuals, before recovering. It is also around this time that the first migrations of modern humans out of Africa would have taken place. The colonization of the Americas by modern humans would have happened between 20,000 and 15,000 years ago, possibly through the Bering strait, not yet opened at that time (Perego *et al.*, 2009).

The real increase in human population took place during the Neolithic Revolution, which occurred around 10,000 years ago with the advent of settlement and the invention of agriculture. The current consensus considers that this discovery happened independently at several times and locations around the world. Some even envision as many as 10 independent discovery of plants cultivation (Diamond, 1999). It seems that bovids may have been domesticated in eastern Sahara before its desertification 7,000 years ago (Kröpelin *et al.*, 2008), which enabled the development of Saharan civilizations, that would move to East Africa and Egypt when the region

began desertifying. Agriculture may have been discovered independently in the highlands of Ethiopia, a view which seems to be strengthened by the botanist Harlan (Harlan, 1969). Harlan describes Ethiopia as a center of diversity of endemic crops that have been cultivated without many changes through centuries, due to the isolation of the Ethiopian highlands.

The domestication of cattle in Africa took place around 10,000 ago (Marshall *et al.*, 2002), leading to the appearance of pastoralists' communities that cohabitated with hunter-gatherers in a large portion of East Africa. It was linked to a need in increased day-to-day predictability, in relation with cultural practices. At first the herds of bovines must have been relatively small, providing maybe only some complement to an otherwise mainly hunted fare. The cultivation of crops probably appeared only later. This dating is important as it indicates that large communities could not have appeared very early in the history of the Horn of Africa. The hunter-gatherers' way of life is not able to sustain large-sized tribes, but mostly small sized groups (less than one hundred individuals).

1.12) Summary : state of the art at the onset of this work.

The work presented here started in 2011. At this time six lineages had been clearly identified in the *M. tuberculosis* complex (Figure 13). This identification had been performed and confirmed using several DNA-based genotyping methods developed during the last twenty years: IS6110 RFLP, spoligotyping, MLVA, partial genome sequencing, whole genome sequencing. Four among these 6 lineages (lineages 1 to 4) correspond to *M. tuberculosis sensu stricto*, whereas the last two lineages (lineages 5 & 6) regroup *M. africanum* and the animal-adapted strains (i.e. *M. caprae*, *M. microti*, *M. pinnipedii*, *M. bovis*). In the *M. tuberculosis* lineages, lineage 1 (a.k.a. EAI) was considered to be the most basal in the evolutionary tree, based on the study of the genomic deletions (Brosch *et al.*, 2002). Indeed the genome of these strains is intact for the TbD1 region as well as the RD9 region, contrary to the other *M. tuberculosis* lineages that are TbD1-deleted. The lineage 2 (a.k.a. Beijing) had been identified as an emerging lineage, particularly linked with the emergence of antibiotic-resistant strains. Due to the study of the deletions and the (admitted) clonality of the MTBC, the long-standing model postulating an emergence of the human-adapted strains from an animal-adapted strain was no more considered as valid, which opened the way to a revision of the major steps in the emergence of this wide-spread pathogen.

Considering "*M. canettii*", it was clear that the strains were more closely related to the MTBC than to any other known mycobacterium. However the exact phylogenetic relationship between the two groups was still subject to discussions. It had been established that even if the isolated smooth strains were pathogenic to humans, they were not transmissible. Therefore this taxon was supposed to evolve in an unknown environmental reservoir. The fact that all strains were found in the Horn of Africa meant that the original location of these strains was probably to be found there. The exact relationship with other environmental bacteria close to this taxon were also largely undefined.

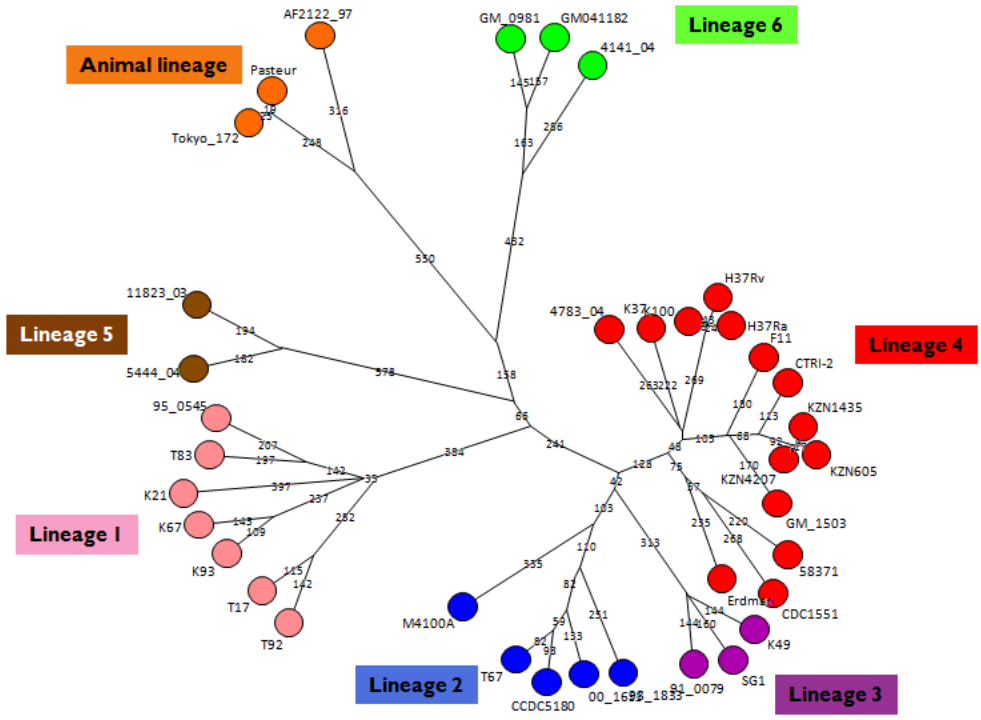


Figure 13: Minimum spanning tree of the MTBC lineages.

Tree obtained from SNPs generated by whole genome alignment of 37 publicly available genomes at the start of this work. The lineages identified by previous typing methods are color-coded on this tree. The number of SNPs is indicated on each branch.

Concerning the probable origin of the *M. tuberculosis* complex it seemed that Africa was the number one candidate for the emergence of this pathogen. However the modalities and the time for this appearance were controversial. In particular, even if it was admitted that the bacterium and its human host should have been co-evolving for quite some time, the spread and differentiation of the extant lineages were not clearly linked with particular migration events in human history.

Considering all those elements it appeared that it would be of interest to characterize in details the genetic diversity of the strains found in the region of the Horn of Africa, both for the MTBC and the "*M. canettii*" strains. Indeed the particular geographic location of the smooth strains seemed to indicate that this region could hold specific clues regarding the evolution of Koch's bacillus, and so was a good target for an in-depth characterization of its genetic diversity, but only if "*M. canettii*" could be confirmed as an outgroup/progenitor species. At the same time the technological developments of sequencing (see next section) made it possible to envision the

sequencing of whole genomes for a given study, enabling the use of whole-genome SNPs as a phylogenetic marker giving a high-level of resolution. Therefore the strategy was to screen a high number of strains by "classical" genotyping methods, and to study in details any particular genotype that would be identified after the global genotyping step. The hypothesis for this work was that some deep-branching lineages of the MBTC phylogenetic tree could be found in Djibouti, providing some indications about the evolution of *M. tuberculosis*.

2) Technical aspects: sequencing and bio-informatics.

2.1) Specificities of high-throughput sequencing data analysis.

2.1.1) Sequencing before the high-throughput revolution.

DNA sequencing appeared after the discovery that DNA was the support of the genetic information in the middle of the twentieth century (Avery *et al.*, 1944; Hershey *et al.*, 1952; Crick, 1954) (Figure 14). DNA sequencing thereafter became a major goal.

One of the most famous of its early developers was Fred Sanger, whose technique, refined during its several decades of use, is still employed today. It was published in 1977 (Sanger *et al.*, 1977), at the same time as the Maxam-Gilbert method (Maxam *et al.*, 1977). This other method was quickly abandoned because it was more technically demanding than the Sanger approach.

The Sanger method is a "sequencing by synthesis" method based on the use of dideoxynucleotides and DNA polymerase. At first, newly synthesized DNA was revealed by autoradiography, thus requiring radioactivity. Subsequently the use of fluorescent labels on the terminating dideoxynucleotides enabled to get rid of the inconvenience of using radioactive elements and most importantly opened the way to automation. Nowadays, the size of newly synthesized fragments is measured using capillary gel electrophoresis to "read" the fragment being sequenced. The length of the sequenced fragments (called reads) can be up to 1200 base pairs, and the average throughput is estimated at 166 kb/hr for a 96-capillary sequencer, according to technical information provided online.

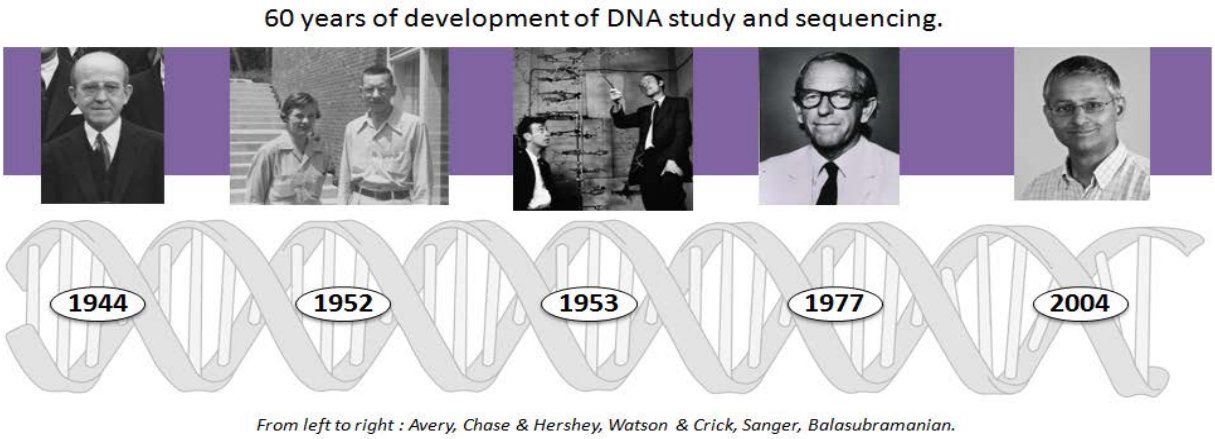


Figure 14: Major figures in DNA research.

Here are presented the portraits of five major figures in DNA structure study and sequencing, with the date of their major contribution to the field.

2.1.2) High-throughput sequencing.

At the start of the twenty-first century new methods fuelled by major prospects in terms of biomedical research arose in the field of genome sequencing. These techniques were developed to facilitate the access to the genomic information encoded in a DNA molecule, with a long-term goal of being able to decipher a human genome at a cost compatible with routine diagnostic needs. Such a goal requires a huge decrease in cost and time necessary to perform the sequencing of a DNA fragment. These goals have been partially reached already during the past 10 years by improving the chemical reactions used and the global throughput of the sequencers (thus the "high-throughput" in the name of these techniques, sometimes also called NGS, for "Next-Generation Sequencing"). The optimization of the throughput is obtained by a massive parallelization of the sequencing reactions.

The principle underlying some of these techniques was pyrosequencing. Pyrosequencing, which is another form of sequencing by synthesis, was invented in Sweden in 1996 (Ronaghi *et al.*, 1996). The detection of the bases is performed using bioluminescence. The read length ranges from 300 to 500 base pairs, shorter than what can be produced using Sanger sequencing. A parallelized version of pyrosequencing was implemented in a machine initially called 454 (from a part of the

name of the company that invented it), which accessed the commercial market in 2004 (Nyrén, 2007). The output of this method is 50 megabases of DNA per hour on average, per machine. This approach is more expensive than the subsequently developed high-throughput sequencing techniques.

The currently most commercially successful approach is "Illumina sequencing", from the name of the company that sells the sequencers (also called Solexa sequencing, name of the original company, subsequently bought by Illumina (Bentley *et al.*, 2008)). The Illumina technology combines both a high throughput and a low cost per base. The read length, which is critical for the reconstruction of the genomic sequence (genome assembly, cf. below), was at first very short. The first Illumina sequencers provided 35 to 50 base pairs reads, to be compared to the 1000 to 1200 bp high quality sequence reads from Sanger sequencing. Today the size of the reads is 300 base pairs and may go up to 400 base pairs in the coming months, due to a improvements of the chemistry of the sequencing reaction. The output of Illumina sequencers is on average around 200 megabases per hour per device.

2.1.3) Challenges associated with NGS data.

Genome sequencing with the current technologies provides fragments (reads) of a sequenced genome. In order to be able to study the genomic information of a given strain, it is therefore necessary to reconstruct the genome from the reads. To use a metaphor, this is like reconstructing a puzzle in order to study the picture that was printed on it. The reconstruction of the genome raises a number of problems, in terms of both hardware and software, and specific answers were developed. The basic difficulty in genome reconstruction results from the presence in most genomes of a variety of repeated elements. For instance the human genome comprises on the order of 500,000 copies of the Alu element, which is an interspersed repeat with a size of 300 bp. Some other classes of interspersed repeats (Kpn elements) are a few kilobases long, and are present in thousands of copies. Another class of repeated elements is tandem repeats (a.k.a. satellites, minisatellites or microsatellites). The heterochromatin part of the human genome which represents a few percent of the

human genome, i.e. hundreds of megabases, is built from short units (a few hundred base-pairs) repeated thousands of times and constitutes "satellite" DNA. As a rule, simple genomes, i.e. viruses, can be reconstituted from relatively short reads. More complex genomes cannot yet be reconstructed from the reads produced by currently available massively parallel sequencing technologies. Bacterial genomes are in between, they can be reconstituted routinely with affordable NGS technologies in a few dozens of contigs.

One issue of NGS is linked to the amount of data generated. The methods mentioned above are "high-throughput", they generate large amounts of reads corresponding to several dozens of time the targeted genomic sequence. Before processing, these reads have to be stored on hard drive disks which necessitates large amounts of storing capacities in the order of the Terabyte (one set of reads for a bacterial genome can be as large as 2 Gb). Nowadays the storage capacity of most personal computers is sufficient to take care of such datasets, but it has nonetheless some impact on the time needed to process such big files (for instance to transfer the data from one computer to another).

Once stored, these short reads have to be used to reconstruct the genome. Two radically different approaches exist, depending on the previous knowledge available: if there is a genome published for the same species, and one wishes only to compare the genomic regions in common with this genome, it is possible to perform what is called reads mapping (or also an homology assembly) (Figure 15). The published genome is used as a reference to re-order the reads. Each read is aligned against the reference genome in order to find its most likely position, which enables to reconstruct the whole target genome. This approach is conditioned by the degree of genetic homogeneity in the studied organism, as only the shared regions between the two genomes will be reconstructed. It is particularly efficient in clonal organisms where all strains are derived from the same ancestral genome, and for the identification of new mutations. If there is no previous knowledge available, or if the genetic heterogeneity is too high for homology to be useful, one has to perform a "*de*

novo assembly", which means that the reads have to be re-ordered without any exterior information. Obviously it is much harder to reconstruct a genome in this way.

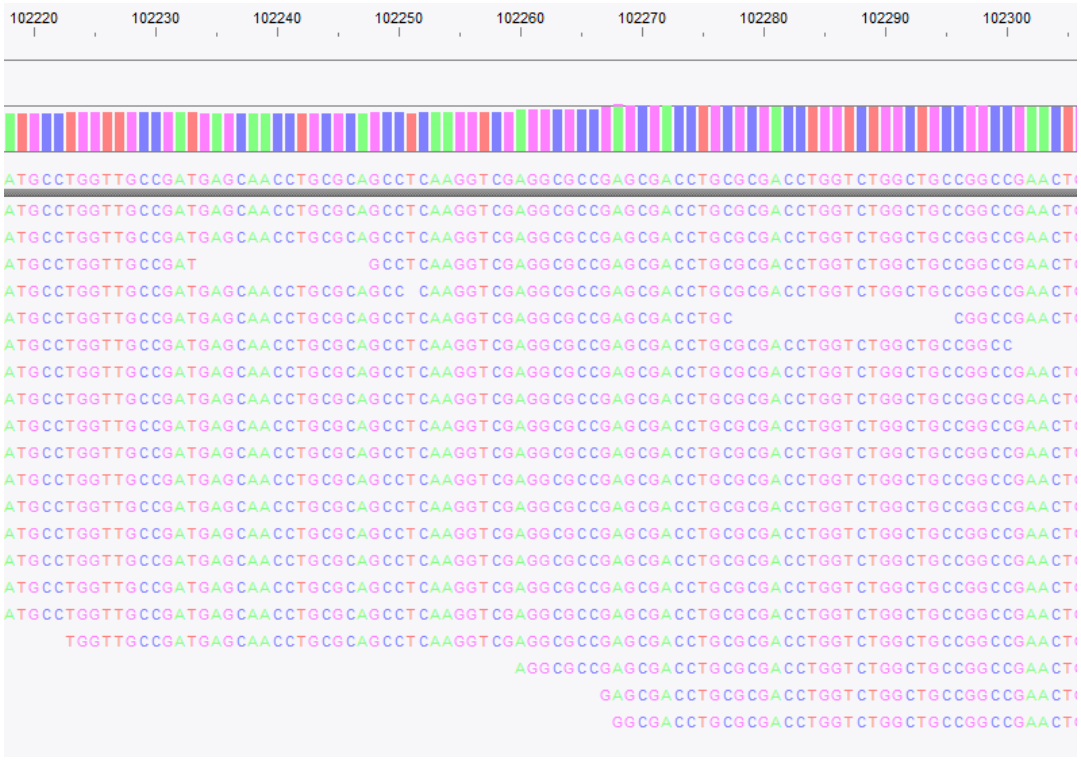


Figure 15: Example of reads mapping on a bacterial genome.

Alignment of reads visualized in the "Power Assembler" module of BioNumerics. Bases are color-coded, the consensus sequence is shown above the gray bar. The vertical histogram is a visualization of the coverage of each base of the sequence. The reads themselves are shown under the gray bar.

For a *de novo* assembly, the similarities between reads are computed to create larger segments made by the concatenation of overlapping reads (called contigs). Different algorithms exist, some based on the use of De Bruijn graphs, as for instance Velvet (Zerbino *et al.*, 2008) and Ray (Boisvert *et al.*, 2010). These two algorithms were optimized for very short reads, and have been adapted to the technological changes. Others, like Mira (Chevreux, 2005), enable the user to mix datasets generated with different technologies, in order to overcome some of the limitations of the reads produced by a given technology. As explained previously, one major problem with short reads is the reconstruction of repeated regions on a given genome. If these repeats show inter-repeat sequence differences close to the sequencing error rate and

are longer than the size of one read, it is difficult to sequence through them, impairing the correct reconstruction of these regions. This results in gaps between the contigs, regions where the exact sequence could not be determined. The number of contigs after a *de novo* assembly of a given bacterial genome, and using the current state of the art of low cost NGS approaches and assembly software, will typically be a few dozens. Strategies such as mate-pair sequencing have been developed to jump over these repeats, and link adjacent contigs (scaffolding) but do not yet provide "closed" genome sequence data. The currently most efficient procedure for bacterial genome sequencing is the combination of Pacific Biosciences (a.k.a. PacBio, wich provides long reads, up to 30kb, but with a relatively low sequence quality) and Illumina data.

Once assembled, either by homology or *de novo*, genomes have to be compared with one another in order to extract useful information from their genomic content. Before whole genome sequencing, epidemiological studies were based on a number of genetic markers that could be investigated by different methods. This included the DR locus (spoligotyping) or VNTRs (MLVA), or, most often for phylogenetic investigation, the partial sequencing of a number of housekeeping genes (Multilocus Sequence Typing or MLST). But these markers, despite their success at describing genetic diversity at the species level, sample only a finite number of relatively small regions of the genome and therefore are not as discriminatory as the data that can be extracted from whole genome sequence comparisons. In addition some of them (spoligotyping, MLVA) have a relatively high homoplasy content. Markers that can be extracted from such whole genome comparisons are the single nucleotide polymorphisms (SNPs). These variations provide an important number of data points that cover a large portion of a given genomic comparison. They are therefore the data of choice for phylogenetic investigations, when one has access to draft genome sequences.

SNPs have to be determined from the genomic sequences being compared. The first step is to align the different genomes and determine the SNPs by looking at the final alignment. This is what is performed by many programs that can produce a genome alignment. It can be useful to know what kind of algorithm was used to

produce this alignment, because sometimes the heuristics can produce artifacts (inducing false SNPs predictions) that will impact the later stages of the analysis.

During these steps of the analysis, one interesting challenge is the size of the generated datasets compared to previous markers. As mentioned about the reads, the transition to using whole genome data has produced an important change in the size of the datasets. In a typical MLVA typing scheme, twenty to forty markers are used to compare a high number of strains, in the order of several hundreds. This produces around ten thousands data points. Using whole genome SNPs, depending on the relationships between the strains one is comparing, there can be thousands of SNPs, times the number of strains compared (from several dozens to one hundred or so in most small to medium scale projects).

2.2) Computing challenges.

2.2.1)Hardware.

As mentioned previously in the description of the data generated by the "next-generation" sequencers, the reads files are large and the hard drive disks capacities have to match this size in order to allow the storage and analysis of various datasets. During the various developments that occurred in the course of this work, more than 3 Terabytes of data files were generated and stored on a desktop computer. This encapsulates both the raw sequenced data (i.e. the reads files), but also the data produced during the different stages of the analyses. One has to take this into account when starting NGS analyses, even if nowadays most personal computers match those storage requirements.

For the assembly process, the software needs to be able to load the data into the memory, and this implies significant RAM needs. Owing to progress in hardware, a slightly improved desk computer is sufficient for this kind of project (I had access to 24 Gb of RAM, a six-cores processor, a SSD hard drive for data analysis and additional hard drives for data storage).

2.2.2)Software

At the beginning of this work Velvet (Zerbino *et al.*, 2008) was the reference tool for *de novo* assembly of reads produced by next-generation sequencers, particularly using the Illumina technology (at this point, reads size was 36, 50 or 75bp). Velvet is based on the use of De Bruijn graphs to compute the overlaps between the reads and generate the sequence of the initial genome. The key element of this algorithm is that it splits reads into k-mers (sub-sequences of size k) and keeps track of the relationship of k-mers from the same read. The k-mers are used to construct the graph and infer the links between the reads in order to reconstruct the sequence. This method cannot take care of large repeats (inherently because of the size of the reads).

Another similar algorithm is Ray, developed by Boisvert *et al.*, that also uses k-mers (Boisvert *et al.*, 2010). Ray is oriented towards parallel processing, and uses heavily the processor capacities of the computer, whereas Velvet needs mostly vast amounts of RAM, which makes a difference in terms of hardware requirements between these two algorithms.

For both approaches the key element is the choice of the k-mer size. As these algorithms implement heuristic approaches for genome reconstruction, it is necessary to test different values of *k* in order to find the optimal one for each project. Most of the time the best *k* value is the one minimizing the number of contigs and at the same time maximizing the total assembled length. Different metrics have been developed in order to be able to evaluate the quality of a given assembly. This problem of heuristic was identified as a bottleneck in data processing, and some solutions were developed, such as Velvet Optimizer, a script running Velvet with different sets of parameters and evaluating the results in order to identify the best one (for a given metric). The time needed to run this type of analysis is the combined time of all the assemblies for the tested range of k.

For different types of data analyses, I have used BioNumerics, a global tool for biological databases management developed by Applied-Maths (Sint-Martens-Latem, Belgium). It is a licensed software (running on Windows), conceived for the analysis and storage of biological experiments. It comprises a number of modules, designed for various applications, as for instance MLVA typing or gel electrophoresis analysis. It can

also be used to analyze phylogenetic markers and draw trees based on those analyses, that can be used to produce publication-quality figures. Its database structure enables to store multiple biologically significant objects, like genomic sequences, MLVA signatures, SNPs...

One strength of BioNumerics is that it is possible to add to the main software some scripts written in Python. They can take advantage of the modules provided by Applied Maths, and tailor their use specifically for the applications needed on a particular project. This enables to develop applications for which the main program may not have been initially intended for.

Why may it be necessary to develop home-made scripts ? The answer lies in the size of the datasets. Data formating steps that could be performed almost manually on small datasets (as for instance for a panel of MLVA markers) have to be automated when one needs to apply them to a NGS-generated dataset.

For instance one of the simplest scripts developed during this thesis project, is a basic function that transposes a table (i.e. inverts rows and columns) contained in a tab-separated text file (script {File_transposition}). A software such as Excel 2010 is able to do it as long as the transposed table width does not exceed 16,384 columns. However some SNPs dataset we encountered were a lot larger, so it was necessary to write such a small snippet of code to be able to manipulate easily the dataset. For a lot of small tasks, it is possible to write a script that will enable to perform them more easily or in a faster way.

A large number of the functions written during this work are related to the treatment of large datasets (at least for a microbiology laboratory), like for instance the merging of two files containing different kinds of information, or some transformations on a file in order to upload it into BioNumerics (which sometimes requires specific file formatting) or other software. Another important part concerned data mining utilities, as for instance a script that enables to look for a specific subsequence (a regular expression pattern) in a given reads-containing file in order to extract all the reads containing this pattern (script {Reads_analysis}). This latest script has also proven very useful in many parts of our projects. It performs a "simple" search

in a file as with a classical text editor. The only difference is that it is not limited by the size of the file considered (it will only take longer to perform if the file is too big), whereas most text editors are not capable of handling reads files larger than 500 Mb.

For a complete list of the scripts written during this work (and their potential use), see descriptive table in Annex C.

II - Results

3) Description of the results obtained during this work and of their background.

3.1) Preliminary considerations. "M. canettii", an outgroup of the MTBC, and the search for unknown M. tuberculosis lineages

In order to try and corroborate the model of the emergence of the *M. tuberculosis* complex in the Horn of Africa, it was necessary to better understand the relationship between the MTBC and the smooth strains that are almost exclusively present in this region. This point was especially emphasized by N.H. Smith (Smith *et al.*, 2009), as it was for instance not yet clear if the "*M. canettii*" taxon could be considered as an outgroup, or was rather a sub-group of the complex, an MTBC lineage which would have gone back to the environment due to contamination from humans or animals.

This question was resolved, in parallel with my own investigations, in early 2013, owing to a work headed by the Pasteur Institute of Lille. Five smooth strains genomes were fully sequenced, and four others were sequenced at draft level (Supply *et al.*, 2013). Among the five completely sequenced genomes of this study, strain STB-D, belonging to the cluster of closely related strains called cluster A by Fabre *et al.* in 2004, was the smooth strain closest to the MTBC in terms of genomic content. This whole genome study permitted to examine genetic transfer events that had been identified previously in "*M. canettii*", as this was one of the elements that interfered in the phylogenetic comparisons with the *M. tuberculosis* complex. The extensive recombination taking place in these strains was confirmed, and shown to have a clear impact on the dN/dS ratio when compared to the MTBC. The transferred segments appeared to be always closer to smooth strains or MTBC genomes than to other mycobacteria. The existence of these transfers was interpreted as supporting the previously proposed environmental origin, possibly in water communities. This environment enables the smooth strains to exchange genetic material, and explain the origin of the infections, as there is not until now any sign that smooth strains are

transmitted from one human to another (Koeck *et al.*, 2011). The authors also analyzed the CRISPRs loci in these genomes and showed that four different types of CRISPRs locus, with different DRs and Cas genes, are present among the smooth strains. One strain (STB-K) was shown to possess two of them. This enabled the authors to identify a list of spacers previously unknown, as only one type of locus (identified as type III-A) was previously investigated by sequencing (van Embden *et al.*, 2000; Fabre *et al.*, 2004; Fabre *et al.*, 2010). The work also presented a biological comparison of the infection by MTBC members or smooth strains, in a mouse model. The 2-times faster growth rate previously determined by van Soolingen *et al.* was confirmed. A lesser persistence of the smooth strains in the lungs of infected mice, as well as a lesser virulence of these strains in the animal model was demonstrated.

These investigations still do not establish the phylogenetic position of the smooth strains compared to the complex. To address this specific point, Interrupted CoDing Sequences (ICDSs) were investigated. These markers had been introduced in 2006 in relation with the evaluation of the quality of a sequencing project, by identifying coding sequences that appeared as interrupted in one organism whereas they were not in another (Perrodou *et al.*, 2006) (Figure 16). The aim of this investigation was to evaluate sequencing error rate in genome sequence data. The authors developed an *in silico* analysis tool that enabled them to identify potential interruptions in a sequenced genome by comparison with the database of all publicly available annotated genomes. Interestingly it appeared after some re-sequencing efforts that in several bacteria these interruptions were genuine and could therefore be used as phylogenetic markers. The analysis was applied to the study of the *M. tuberculosis* complex, where 81 ICDSs were predicted as being characteristic of all MTBC members (i.e. interrupted in the complex and intact in all other closely related organisms) (Deshayes *et al.*, 2008). It thus seemed possible to use ICDSs to establish if the "*M. canettii*" taxon could be considered as an outgroup relative to the MTBC (Smith, 2006). Since the "*M. canettii*" genome data was not yet available in 2008, the 81 ICDS prediction was made independently of "*M. canettii*". The ICDS status in "*M. canettii*" was analyzed by Supply and col., who found that four were not interrupted in

all their "*M. canettii*" strains. This definitely established the outgroup status of "*M. canettii*".

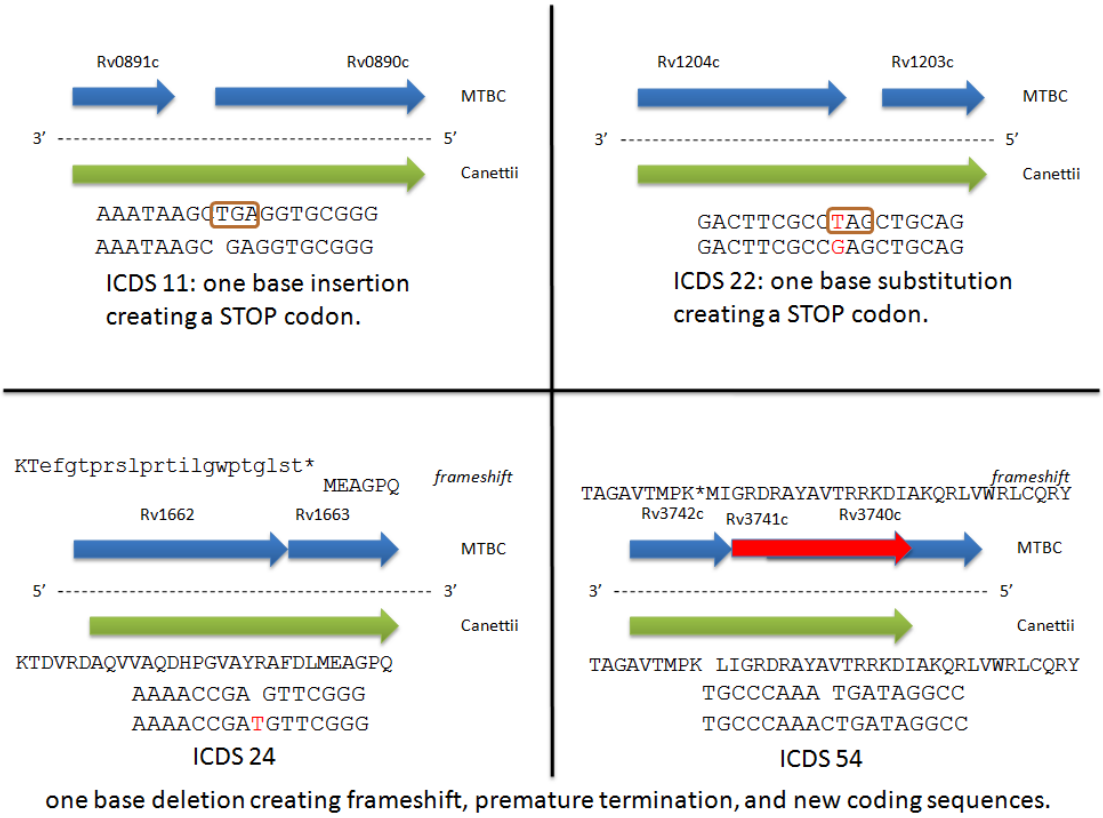


Figure 16: Details of the four ICDSs differentiating "*M. canettii*" strains from the MTBC members.

In each case the nucleotide sequences are shown, with an emphasis on the mutation causing the proteinic difference between the two species. For the two frameshifts the amino-acids sequence is also displayed.

We investigated the 81 ICDS on the extended SSA collection of "*M. canettii*" including strains not analyzed by Supply and col. (Figure 17). All the strains investigated show the same differences between *M. tuberculosis* and "*M. canettii*" for the four ICDS identified by Supply *et al.* The interruption in members of the MTBC appeared to be caused by a single point mutation that could be identified in comparison with all smooth strains (Figure 16). All the "*M. canettii*" strains showed the same sequence at these positions, in the absence of transfer events, showing that the absence of interruption was a shared trait that had to be present in the common ancestor of all the smooth strains. Curiously the majority of the investigated "*M. canettii*" appeared

to show the same status in terms of ICDS, with no other non-interrupted element out of the 81 characteristic of the MTBC members. A notable exception is shown in Figure 16 for ICDS 38 in Percy 302 / STB-K. This strain is the most divergent of all the smooth strains and seem to possess an extra not interrupted region among the 81 ICDS. This peculiarity excepted, the observed homogeneity in terms of ICDSs in the "*M. canettii*" strains is quite unexpected, as one could have anticipated that several "*M. canettii*" lineages would be positioned at different places along this ladder of 81 phylogenetic steps. It also shows that even if the smooth strains are genetically more diverse than the strains of the *M. tuberculosis* complex, they share nonetheless a high degree of genetic homogeneity.

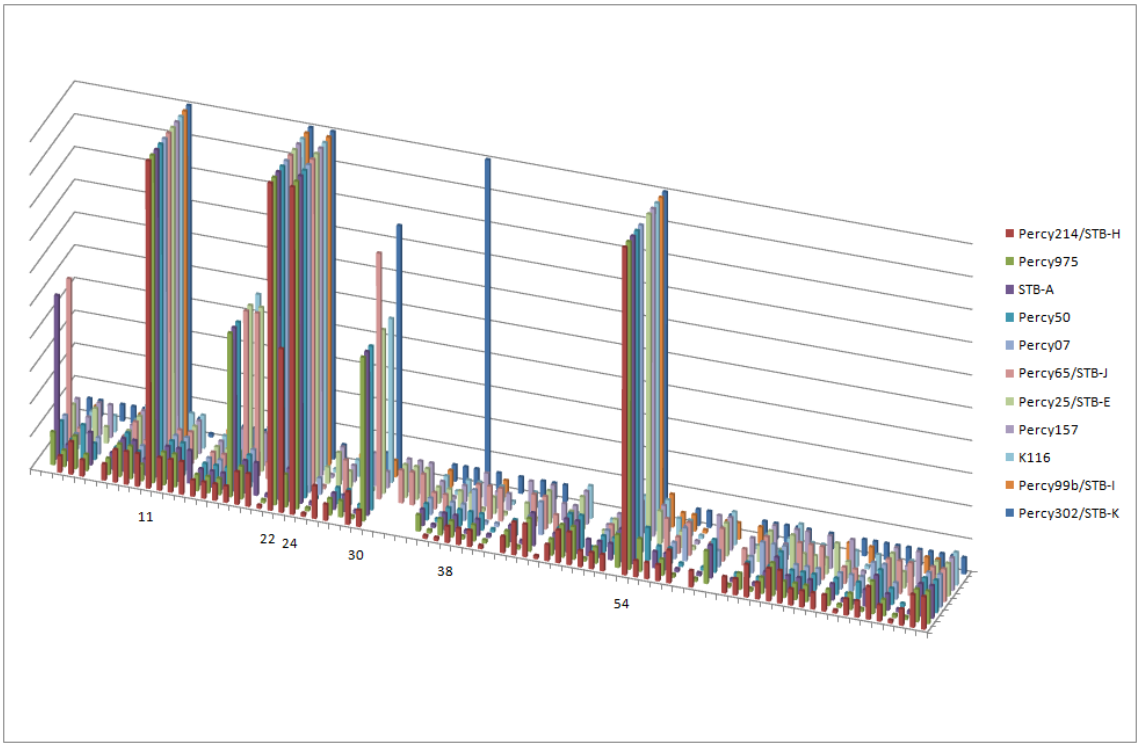


Figure 17: Evaluation of the interrupted status of the 81 ICDSs on several strains from our collection.

Results from the *in silico* analysis of the high-throughput sequencing reads of strains from our collection. Full-sized peak: no interruption, smaller ones: interruption as in the MTBC, intermediaries: varying situation. White spaces correspond to missing data, when the ICDS region could not be reconstructed from the NGS data. The strains whose name does not contain "STB-.." were not investigated in (Supply *et al.*, 2013).

The formal confirmation of the outgroup status for "*M. canettii*", making "*M. canettii*" the modern representatives of the progenitor species from which *M. tuberculosis* emerged, as speculated before (Fabre *et al.*, 2004; Gutierrez *et al.*), and its geographic localization provided strong support to an East-African origin for *M. tuberculosis*. The next step was to characterize precisely the genetic diversity encountered in the Horn of Africa (cf. map Figure 9). If the "Horn-of-Africa" origin was correct, it was expected that there should be a high diversity of *M. tuberculosis* lineages, and maybe even the presence of unknown lineages branching outside of the known MTBC tree (i.e. lineages that could pre-date the MRCA of the contemporary lineages). In conjunction with the study of smooth strains, MTBC strains have also been isolated by the French Army Health Service operating in the Republic of Djibouti, leading to the constitution of an extensive collection of strains from this part of the world. We took advantage of this collection to test this prediction of an East-African origin of the MTBC.

In Article I, the MTBC collection from Djibouti lead to the identification of an exceptional lineage. This lineage is unique because it shows a very limited geographic distribution (the Horn of Africa) but is also a very deep branching lineage. We were surprised that this lineage does not appear as an outgroup for the other six lineages. Article I provides an interpretation of this finding. In article II, the ongoing outbreak of "*M. canettii*" infections is investigated and used to further root the MTBC complex. The datasets used in these two studies can be combined in order to root the tree obtained from the MTBC strains, as the smooth strains belonging to clone A and responsible for the current outbreak were shown to be the closest outgroup of the *M. tuberculosis* complex (i.e. they were genetically the closest, in terms of shared ancestry, while at the same time belonging to another genetic group). To perform this analysis, the dataset containing the Djibouti MTBC strains and additional genome sequence data was mixed with the clone A genomes. All sequences have been assembled using *M. tuberculosis* H37Rv as a reference in order to generate an homogenous dataset. This enabled to determine the SNPs for this new dataset and to draw a new MST. It also allowed to root precisely the MTBC tree and to propose a detailed evolutionary

scenario (developed in the first article). We hypothesized that the newly discovered lineage 7 (presented in the first article) constitutes the local ecotype of *M. tuberculosis,* that remained in its original environment, whereas the other lineages have diverged in new ecotypes due to their expansion outside the Horn of Africa.

3.2) Article I: Significance of the Identification in the Horn of Africa of an Exceptionally Deep Branching Mycobacterium tuberculosis Clade.

Summary:

Based on the isolation of the smooth strains only in the Horn of Africa, this work was aimed at characterizing extensively the genetic diversity of the *M. tuberculosis* strains found in the same region, in the hope of finding evidence that would confirm the model of an East-African emergence of this pathogen. A collection of more than 400 strains collected over a decade was analyzed using MLVA and spoligotyping, and NGS sequencing was used for the in-depth study of some peculiar strains. This study reveals the most detailed picture to date of the genetic diversity of the *M. tuberculosis* strains isolated in the Republic of Djibouti, and may provide an indirect estimate of the overall diversity in the whole Horn of Africa, as the Republic of Djibouti is a place of extensive population mixing. Five of the six lineages (Africanum lineage 5 was not detected) of the MTBC were identified, illustrating the important diversity observed in this part of the world, even if Africanum lineage 6 was represented by only two strains. Whole genome sequencing of strains identified as peculiar by genotyping confirmed that they corresponded to unusual branches in the MTBC phylogenetic tree. Moreover, this exhaustive characterization lead to the discovery of a new lineage, named lineage 7. The interpretation of lineage 7 position in the MTBC leads to the new model of evolution proposed for the complex in this thesis work.

Link to the article: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531362/

3.3) Article II: Progenitor "Mycobacterium canettii" Clone Responsible for Lymph Node Tuberculosis Epidemic, Djibouti.

Summary:

This article investigates smooth strains from recent outbreak events, in order to characterize them at the genomic level, and elucidate their relationship with the MTBC. The object of the study are the smooth strains isolated in the region of Djibouti, and responsible for an outbreak of tuberculosis in adults and children. In this study the previously mentioned cluster A is shown to be one particular emerging clone responsible for the majority of recent cases. Its characterization leads to the identification of one horizontal gene transfer event. As this clone regroups the smooth strains most closely related to the *M. tuberculosis* complex, its study enables also to root precisely the MTBC phylogenetic tree. According to this branching, an evolutionary model can be proposed, together with some elements to evaluate the age of the MRCA of "*M. canettii*", in conjunction with the results obtained from the study presented in the first article. This leads to the proposition of a scenario, postulating the emergence of the MTBC in the Horn of Africa from an environmental bacteria that was probably closely related to the MRCA of the "*M. canettii*" strains.

Link to the article: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3884719/

III - Discussion

4) Emergence and evolution of the MTBC.

4.1) Model of emergence.

The precise rooting of the "*M. canettii*" outgroup and the proposed interpretation of lineage 7 lead to a refined emergence model for the *M. tuberculosis* complex. The complex is the remnant of two "out-of-the-Horn-of-Africa" waves. The first is represented by superlineages one and two. The second corresponds to the third superlineage. This new model predicts that key evolutionary events occurred between the departure of the first and second superlineages on one hand, and the third on the other hand. We hypothesize that the finding of lineage 7 will allow to reconstruct *in silico* the most recent common ancestor of these key branching points, and help identify candidate genes responsible for the successful spreading of superlineage 3 and replacement of superlineage 1. Thus the analysis of parameters that could influence the diversity of bacterial strains, and the estimation of time scales, can bring new insights for the construction of strategies to fight these infectious agents.

Because of their high genetic proximity with the MTBC, the study of the smooth strains, the "*M. canettii*" taxon, is of particular interest. This opportunistic pathogen causes diseases with the same clinical signs as infections caused by members of the MTBC, without so far any evidence of inter-humans transmissions. Our study of the recent outbreak events of lymph nodes infections in children in Djibouti has shown the clonality of the group of strains causing these infections. As these strains are forming the majority (more than 70%) of the "*M. canettii*" isolated during the last decades, it seems likely that clone A is an emerging pathogen in this region of the world. Due to the high similarity with *M. tuberculosis* we postulate that this emergence may mimic the mechanism that has enabled the emergence of the MTBC, as well as the acquisition of its current characteristics. The precise determination of the prevalence of "*M. canettii*" infections in the Republic of Djibouti is made difficult because of likely sampling biases, linked with the fact that they have been collected in French military hospitals. The statistical differences between the isolation rates from expatriates and

from local patients suggest that the prevalence of "*M. canettii*" infections in Djiboutian children is underestimated. Another hypothesis is that Djiboutian encounter "*M. canettii*" early in their life, possibly in an environmental reservoir, and become immunized. Given the infections caused by other closely-related mycobacteria, the initial form of an infection by *M. tuberculosis* in the early stages of its evolution may have been a lymphadenitis, thereafter evolving towards a pulmonary form, transmissible from human to human. This characteristic would be acquired secondarily during the co-evolution between the pathogen and its host (Behr, 2013).

In order to better understand the mechanisms underlying the emergence of pathogenicity, it would be necessary to determine the environmental reservoir of the smooth strains. The existence of such a reservoir is supported first by the absence of inter-humans transmissibility (for this ecotype, humans represent only a "spillover host" and not a maintenance host (Smith *et al.*, 2009)), and then by the existence of recent horizontal gene transfer events (see Results) that are possible only in an environmental setting, because of the intra-cellular multiplication of this pathogen in humans. The identification of spacers corresponding to *M. marinum* prophages in the CRISPR locus of some smooth strains is pointing towards an aquatic reservoir, possibly in relation with amoebae (Mba Medie *et al.*, 2011). However until now no such source has been confirmed. It is notable that the duration of the contact before infection seems to be rather reduced, as one of the infected patients from the clone A study (a child) had been present for only four months in Djibouti before the identification of the infection. Investigations have to be continued in order to identify the origin of the smooth strains, and thus refine the knowledge of the acquisition of pathogenicity, as well as of the factors controlling HGT events. Up to these days, the origin of the exogenous fragments identified in the smooth strains could not be determined. These genomic fragments come from an organism which may show well over a 1% divergence with the smooth strains, more than an order of magnitude higher than the divergence observed between two "*M. canettii*" strains (or the one observed inside the MTBC). This organism, probably living in the environment, remains unknown at present. Besides these investigations, the surveillance of the cases caused by clone A

members must be pursued. In the hypothesis where one strain could acquire the capacity to be transmissible between humans, the identification of the genetic mechanisms causing this change could shed light on this trait in *M. tuberculosis*, opening new opportunities for treatment (without mentioning that early detection of such strains could permit to limit propagation).

The detailed study of clone A strains (*via* high-throughput sequencing) has enabled to analyze the number of mutations in relation to the date of isolation of the different strains. The results show a similarity between the number of mutations differentiating "*M. canettii*" strains from the ongoing outbreak, and those evaluated in the diverse studies that have explored outbreaks (Gardy *et al.*, 2011; Walker *et al.*, 2013). The number of SNPs by branch in clone A is of the same order of magnitude than in an MTBC outbreak. This suggests that, despite differences, for instance in terms of transmissibility, "*M. canettii*" evolves at the same rhythm or more slowly than *M. tuberculosis*, but not faster. This could be linked to the biological mechanisms involved in the resistance to the degradation inside macrophages for both bacteria.

Finally, one other interest in the study of the clone A resides in its proximity with the *M. tuberculosis* complex, as compared to other "*M. canettii*". As suggested by genotyping (Fabre *et al.*, 2010), and later confirmed by whole genome sequencing (Supply *et al.*, 2013), clone A regroups the "*M. canettii*" strains closest to the MTBC in genetic terms, even if they belong to a distinct phenotypic group, as defined by the analysis of the ICDSs. This proximity is not necessarily reflecting a more recent common ancestor as compared with other "*M. canettii*" lineages. Rather it essentially reflects a larger proportion of genome not affected by HGT events.

This leads to the question of the definition of a "core genome" for "*M. canettii*" in order to determine its "real" genetic distance to the complex when all transfer events are filtered. This question could be solved by looking at the composition of the smooth strains' genome in terms of ICDSs. It appears that all "*M. canettii*" strains share the same four non-interrupted ICDSs, with one exception (Figure 17). This seems to hint to a high homogeneity of the genome when transfer events are not taken into account, suggesting that all the observed smooth strains may descend from a common

ancestor which would have possessed the same genomic content for these 81 genomic markers. Figure 18 summarizes our current view of the "*M. canettii*" ecotype evolution. In this view, all the "*M. canettii*" strains sampled in the Horn of Africa represent a unique ecotype. Clone A is currently predominant and will eventually replace the others. The current MTBC represents a "frozen" "*M. canettii*" sampled out of this ecotype to colonize other environmental niches, leading to new ecotypes. From this sampling event to the tips of the current MTBC lineages, about 23 ICDSs occurred according to (Deshayes *et al.*, 2008). In this regard it would be interesting to investigate ICDSs which occurred along the "*M. canettii*" lineages. Due to the sampling bias and the unknown environmental reservoir of these strains it is however difficult to get a clear picture of the genetic diversity that one may observe in this particular ecotype. Clone A constitutes therefore the most appropriate outgroup to root the MTBC with respect to our current knowledge of this diversity.

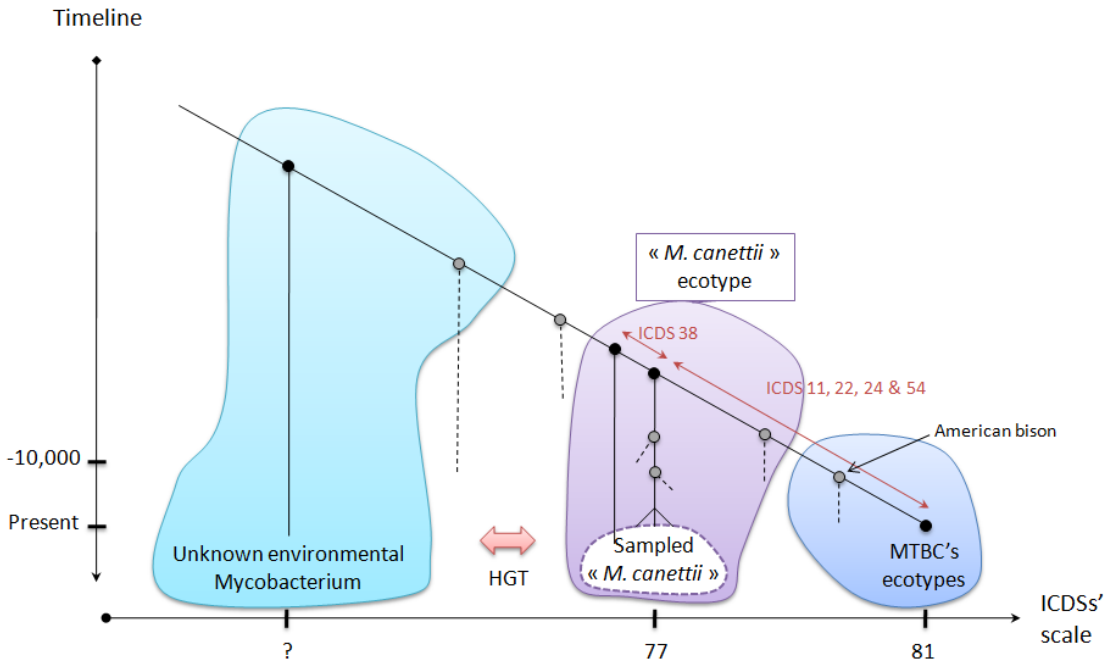


Figure 18: Diagram of the proposed population structure and evolutionary history of "*M. canettii*" and the MTBC.

This diagram combines a potential timeline with the ICDSs' differences observed among the 81 ICDSs characteristic of the MTBC. Each colored surface correspond to an ecotype seen in time, and centered on its ICDSs' status.

This brings us to the results of the study of the *M. tuberculosis* strains from Djibouti. As mentioned before, the choice of this particular location was made according to one key element, the geographical restriction of the smooth strains to the Horn of Africa.

A first lesson from this analysis resides in the use that can be made of classical genotyping techniques, despite the ever increasing developments of high-throughput sequencing technologies. Spoligotyping and MLVA typing remain much less costly and generally faster, which enables to run a preliminary screen of a large number of samples, in order to identify the most interesting strains. The sequencing of a collection of more than 400 strains could not be envisioned because of the costs of such a high-risk project. Spoligotyping alone would not have been sufficient. Although spoligotyping has proved very efficient for the global description of the MTBC complex, it does not provide reliable distance estimates, and is not robust to recognize rare and deep branching lineages such as lineage 7. This is illustrated by the finding of lineage 7 strains in the SITVIT Web database (Blouin *et al.*, 2012). The combination of these two approaches proved fruitful. The genotypes identified as peculiar did correspond to very particular strains.

The analysis of the diversity of the *M. tuberculosis* strains in Djibouti enabled to identify strains branching very internally inside the MTBC tree. It could be correlated with an origin of the complex from this region of the globe, as these internally branching segments have probably diverged early during the first steps of the diversification of the MTBC. The most striking examples are the two strains defining lineage 7. Given the starting hypotheses for this project (African origin of the complex and location of "*M. canettii*"), it was anticipated that particular strains of the *M. tuberculosis* complex could be identified, representing ancestral branching points, and bringing new insight on the emergence of the MTBC and the different steps of its evolution. We expected that these ancestral branching point would constitute outgroup lineages with respect to the currently known MTBC (Figure 19). This would have corresponded to a single departure from the Horn of Africa. Whole genome sequencing and the resulting MTBC topology showed that the history of the early

emergence of the complex includes two departures from the Horn of Africa leading to all extant lineages.

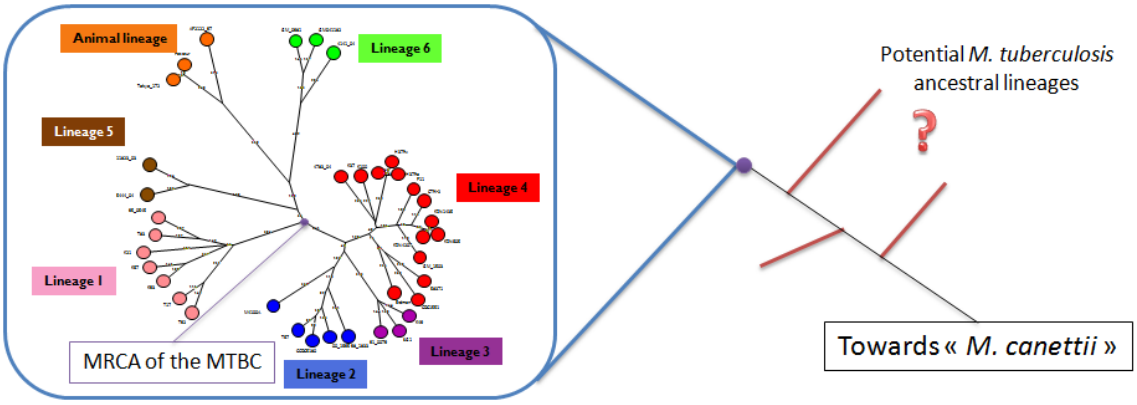


Figure 19: Hypotheses regarding extant "ancestral" lineages of *M. tuberculosis* before the Djibouti study.

This diagram illustrates the hypotheses that could be envisioned before the Djibouti study regarding the departure of ancestral lineages: such lineages would branch outside the tree of the extant lineages of the MTBC.

Lineage 7 is particularly remarkable because of its rarity, very tight geographic distribution, very deep branching position in the MTBC, explaining why it had not been described before our targeted investigation of *M. tuberculosis* in Djibouti. In another study performed almost simultaneously in Ethiopia, Firdessa and col. reported a remarkable prevalence (approximately 15%) for this lineage in North-East Ethiopia, near the Republic of Djibouti (Firdessa *et al.*, 2013). It may be limited to this geographic area, as only a few strains from Kenya showed the lineage 7 signature in SitVitWeb (Demay *et al.*, 2012) (presence of spacers 34 and 39). The proportion of lineage 7 in the Republic of Djibouti is similarly very low, and may be associated with Ethiopian patients. The notion of bacterial ecotype provides a very helpful conceptual framework to interpret lineage 7. Inside the *M. tuberculosis* complex all the strains are genetically very homogenous (with a mean genetic divergence around 0.05%). Consequently it may seem questionable to assign species level to some lineages inside the MTBC, following the binomial nomenclature (as for instance *M. bovis* or *M. africanum*). These species correspond to the definition of ecotypes. The animal lineages strains may have been protected from the subsequent spread of superlineage

3 precisely because they had enough time to adapt to new hosts. *M. africanum* would have been preserved, may be temporarily, and only where protected by relative geographic isolation (Western Africa). In this acceptation, lineage 7 is a very peculiar ecotype, restricted to a precise area of the world (i.e. a "geotype"), and this has a major impact as this location is linked with the supposed origin of the *M. tuberculosis* complex in Africa. It is then possible to hypothesize that lineage 7 corresponds to the local ecotype marking the region of emergence of the MTBC, and from where it has diversified. This would predict that no other unknown deep-branching lineage from the *M. tuberculosis* complex remains to be found. The identification of lineage 7 was only made possible because of the sufficient geographic proximity between the sampling region (i.e. Djibouti) and the current maintenance locus of this bacterial population, which may be Ethiopia (Firdessa *et al.*, 2013).

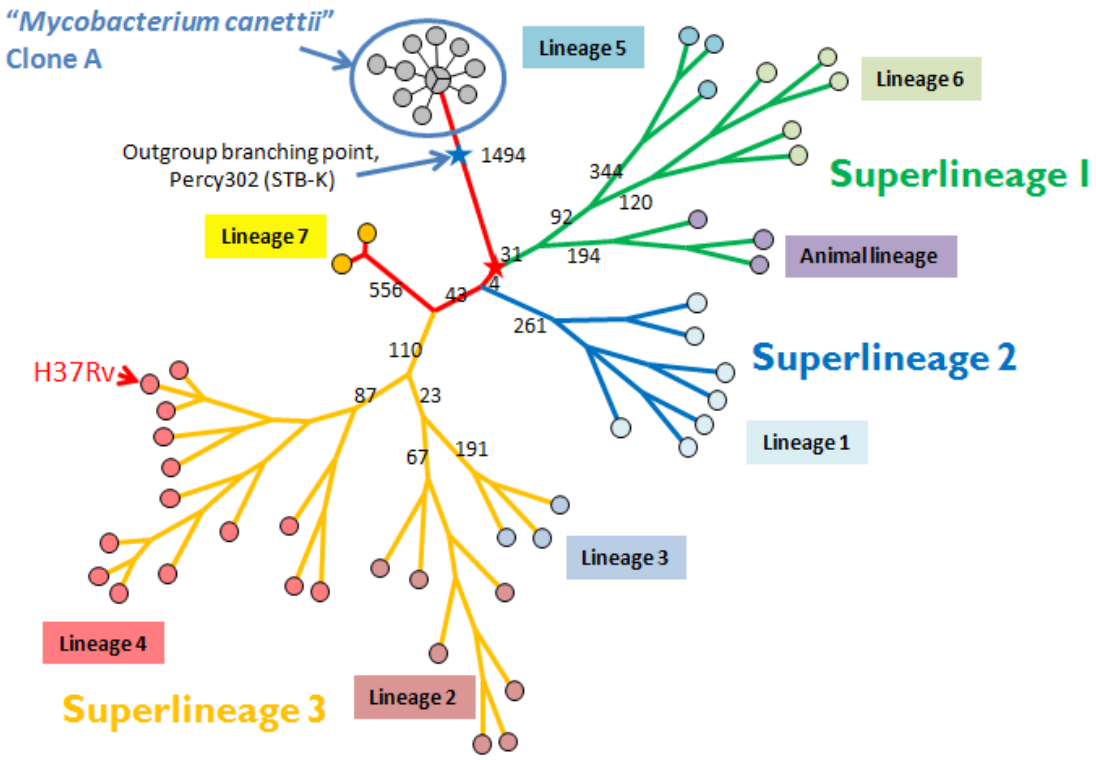


Figure 20: Phylogenetic tree of the superlineages of the MTBC.

Modification of figure 2 from (Blouin *et al.*, 2014). Global phylogenetic tree showing the three main superlineages of the MTBC (green, blue and yellow branches) and "*M. canettii*" clone A (gray nodes). The path leading from "*M. canettii*" to lineage 7 is shown in red. The "classical" lineages of the MTBC are color-coded in the nodes.

It is possible to envision a new model giving more importance to these considerations. As "*M. canettii*" is an outgroup to the complex while being at the same time pathogenic but non-transmissible from human to human, one can suppose that the ancestor of the complex is strongly related with the ancestor of "*M. canettii*". The co-localization of the smooth strains and the lineage 7 strains in the region of the Horn of Africa, as well as the genomic characteristics of this lineage, permit to suppose that it represents the contemporary result of the evolution of the local ecotype from where all the other current lineages of the MTBC diverged. The separation of the different lineages along a linear tree leading from "*M. canettii*" to lineage 7 (Figure 20, Figure 21) suggests the emergence of three super-lineages that all departed from the Horn of Africa. A first exodus would have led to the West-African and the animal-adapted lineages (lineages 5 & 6), and four SNPs later the departure of lineage 1 towards the Indian Ocean and the South-East of Asia. The departure of the lineages that are sometimes called "modern" (lineages 2, 3 & 4) would have taken place in a second time, as the divergence of these lineages is more "recent" in terms of the number of SNPs that separates them from the previous splits.

Some questions are still opened. Whereas lineage 7 appears to be mainly restricted to the highlands of North-East Ethiopia, "*M. canettii*" is most frequent in the Republic of Djibouti. The expected geographic coincidence is not fully satisfied. More investigations will be needed to better pinpoint the cradle of the MTBC. It may be that lineage 7 originates from the Republic of Djibouti, but was preserved in neighbouring Ethiopian highlands owing to the isolation of this region. Or on the contrary, it may be that "*M. canettii*" is present also in the Ethiopian highlands but has not been detected there yet, because of a lower medical surveillance, a lack of naïve population or a lower development of the water network.

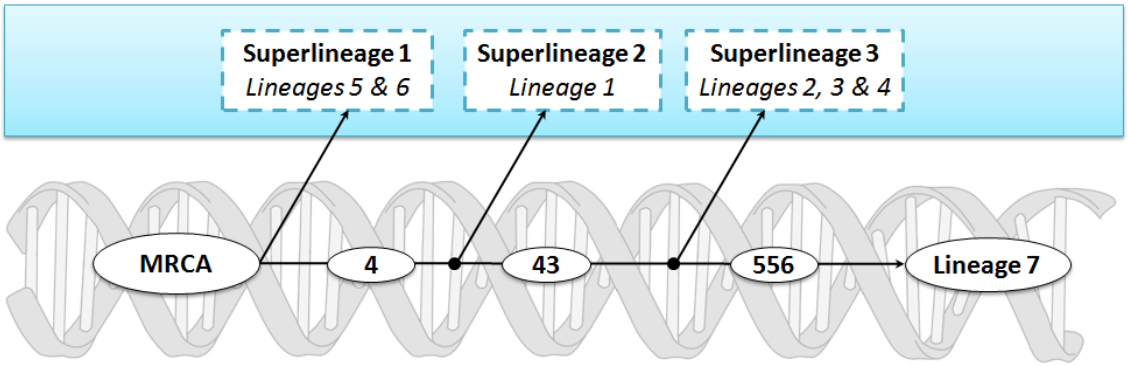


Figure 21: Linear model of evolution for the M. tuberculosis complex.

Linear tree from the MRCA of the MTBC as identified by the branching point of "*M. canettii*" towards *M. tuberculosis* lineage 7, with the projection of the three super-lineages (projection on the red path of the previous tree). On the axis, number of SNPs separating the different events.

4.2) Dating the model.

In order to try to infer a relationship between the observed number of SNPs and time points in human history, it is necessary to present the elements that could help establish a link between the observed mutations and the evolutionary history of the pathogen. The key element for this understanding is the notion of substitution rate, *i.e.* the number of mutations that are fixed on average in a genome during a certain period of time. If this rate could be determined then it would be possible to date the divergence between lineages, by looking at the number of single nucleotide polymorphisms differentiating them. According to the macaque model studied by Ford *et al.* (Ford *et al.*, 2011), the mutation rate per genome per day for *M. tuberculosis* is remarkably constant (whatever be the state of the disease, active or latent), around 0.1 mutation per genome per year. This estimate provides an age of about 10,000 years for the MRCA of the complex. Figure 20 and Article II then indicate a maximum age for *M. tuberculosis* itself. The branch length from the MRCA of "*M. canettii*" clone A and *M. tuberculosis* to the MTBC MRCA is approximately 1500 SNPs. Percy302 (STB-K), which can be considered as an outgroup according to the ICDSs investigation, branches at a distance of apporximately 1000 SNPs from the MTBC's MRCA and 500 SNPs from clone A. The length from the MTBC MRCA to the tip of lineage 7 is approximately 600 SNPs. This suggests that the substitution rate in the *M. tuberculosis*

ecotype is 3-times faster compared to the "*M. canettii*" ecotype (1600 versus 500 SNPs). Finally, if the MTBC MRCA is 10 kY old (600 SNPs), then *M. tuberculosis* cannot be more than 25 kY old (less than 1600 SNPs).

This represents the lower range of estimates in terms of MTBC dating. Since the last twenty years two models coexist, because of the weakness of experimental evidence (such as studies of ancient DNA) : the "young MTBC" model (the MTBC MRCA would be 10-20 kY old (Sreevatsan *et al.*, 1997)) and the "older MTBC" model (the MRCA would be 40 kY or more). The comparison of these models is linked to the extent of co-evolution between the pathogen and its host. For those who defend the oldest date, there has been co-evolution on a long-time scale between Man and bacteria, which led to an important adaptation of the later to the different human phenotypes. When humans migrated out of Africa, they were already infected by *M. tuberculosis*. On the contrary for those who defend a more recent emergence, the link between human phylogeography and that of the complex certainly reflects a degree of adaptation but does not prove the existence of a strong co-evolutionary linkage. *M. tuberculosis* spread later, along commercial routes.

One of the most recent reports on this topic supports the older MTBC model. According to the authors, the MTBC emerged in Africa 70,000 years ago (Comas *et al.*, 2013). In order to underpin this model, the authors have chosen to discuss several aspects. They tried first to determine the most probable geographic region for the emergence of the MTBC, using statistical tests based on the distribution of the different lineages. This first step designated Africa, and more precisely the Eastern part of Africa, as the most likely region for the emergence of the MRCA of the complex. This step was essentially a statistical validation of the models previously developed on the emergence of the MTBC (cf. introduction). Then the authors attempted to settle between the different time points presented in the literature to date the emergence of the complex. In this article the compared temporal models were 185 kY, 70 kY (and a 65 kY variant) or 10 kY.

In order to address this question, Comas *et al.* compared a phylogenetic tree generated from the analysis of human mitochondrial haplogroups with that of the

complex. They claimed to find a striking coincidence and congruence between both phylogenies, and then they ruled out the use of short-term substitution rates determined for the MTBC (Ford *et al.*, 2011) to extrapolate longer term rates. Their objection was based on a study from Morelli and col. (Morelli *et al.*, 2010) establishing that, in *Helicobacter pylori,* the long-term mutation rate was 5 to 17 fold lower than the short term one. Consequently if the estimated mutation rate is considered as inadequate, this leaves only the use of statistical models to evaluate the likelihood of the different proposed models.

Therefore with the help of this mathematical method and by trying to establish links between the major time points in human development and the diversification of the different lineages inside *M. tuberculosis,* the authors proposed an emergence of the complex 70,000 years ago in East Africa. The diversification occurred with time steps of several thousand years, accompanying the migratory moves of early modern human populations. The first were towards West Africa around 73,000 years ago, then towards the Indian Ocean 6,000 years later or towards Europe and Central Asia 46,000 years ago. Despite this very ancient dating of the major splits between the different lineages, the authors detail also the importance of the Neolithic Demographic Transition (NDT) (starting around 10,000 years ago) for the diversification of the complex. The augmentation of the size of the human population induced the augmentation of the effective population size of the pathogen.

Although at first glance the model looks convincing, a more detailed reading points to some highly speculative points. Firstly, the "*M. canettii*" taxon is never mentioned in this work, and this despite the notable genetic proximity between it and the MTBC. This is surprising because the exclusive presence of "*M. canettii*" in the Horn of Africa is an element that rather strengthens the hypothesis of an emergence in East Africa, without the need for any sophisticated models. Along the same line, the lineage 7 is considered by the authors as simply another MTBC lineage, "lost" in an isolated location after being separated from the rest of the complex, some 64,000 years ago. This view of lineage 7 is essential to support the claimed resemblance between the phylogeny of the human mitochondrial genome and the MTBC

phylogeny. However, there is no such resemblance if lineage 7 is viewed as the "original ecotype" emerged from "*M. canettii*". The lineage 7 ecotype model that we published in 2012 is not discussed by Comas *et al.* Secondly, the author's rejection of the extrapolation of short term substitution rate to long term substitution rate is not supported in the case of the MTBC complex by any significant variation in the dN/dS ratio observed among the branches of the MTBC tree (Figure 22). Indeed in the *H. pilori* investigation by Morelli and col., the difference is mainly explained by the effects of purifying selection in this bacterium, and translates in the evolution of the dN/dS ratio with time. Consequently, the two main arguments proposed by Comas *et al.* appear to be very weak at best. The other arguments are circumstantial, as it is for instance demonstrated in (Pepperell *et al.*, 2013). According to this study there is no clear signs of co-divergence that would link the phylogenetic structure of the MTBC with that of the human population. Moreover, the results of the *in silico* analyses developed in this article lead to the determination of a mutation rate that is perfectly compatible with the one determined by the macaque model presented in (Ford *et al.*, 2011), even concluding that this rate does not seem to change with the time-scale of the sampling, contrary to what is stated by Comas and col..

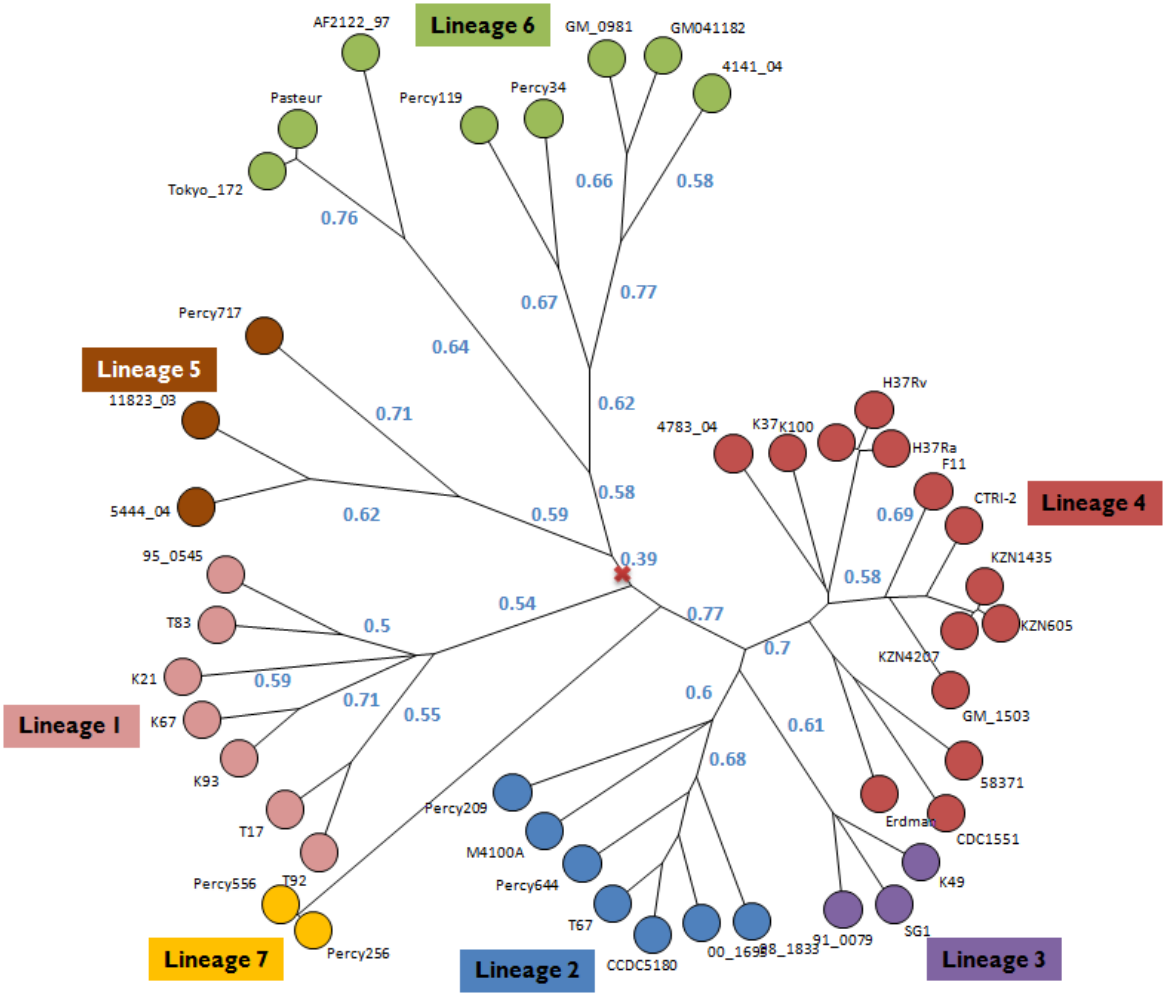


Figure 22: Ratio of non-synonymous to synonymous variations within the MTBC.

The MRCA position is indicated by the red cross, the ratio is displayed in blue next to each branch for which it was determined. The lineages of the MTBC are color-coded in the node corresponding to each genomic sequence.

It is possible to make a number of alternative hypotheses regarding the dating of the *M. tuberculosis* complex. Together with the "older MTBC" emergence models, it was proposed that, due to the specific characteristics of tuberculosis (such as its capacity to stay in a latent form in infected individuals), it had to have appeared before the rise in the population densities, and so before the Neolithic Demographic Transition. However some elements are compatible with another interpretation, in the light of the behavior of contemporary strains of *M. tuberculosis,* but also of "*M. canettii*". First, one can analyze the prevalence of tuberculosis in contemporaneous populations of hunter-gatherers, whose way of life is likely very close to the one of our

pre-Neolithic ancestors. A fundamental characteristic of these human groups is their small size, in average around 100 individuals, if seasonal groupings are excluded (Hamilton *et al.*, 2007). It is probable that a reduced size was also a characteristic of the first modern human groups that have progressively migrated outside of the African birth ground, during the early steps of the human spread (Liu *et al.*, 2006). A numerical simulation has established that the minimal population size, in order for tuberculosis to be maintained as an endemic infectious disease, was between 180 and 440 individuals, depending on the population models used (but postulating a hunter-gatherer environment) (McGrath, 1988). This size is certainly larger than that of the first human groups in the model proposed by Comas and col.. Several studies on current population of hunter-gatherers confirm that these populations do not suffer from the same epidemics as populations living in settlements, even if they belong to the same genetic groups. Besides, some ethnologic studies have established that today tuberculosis is not endemic in hunter-gatherers, and is found only when they are in contact with a settlement where it could be maintained (Gurven *et al.*, 2007). It is also worth mentioning that the Australian Aboriginal population did not suffer from tuberculosis before the colonization (Burns, 2003). This element goes against the hypothesis that every migratory groups have contributed to the world-wide spread of tuberculosis, as postulated in the co-evolution model proposed by Comas *et al.*.

Consequently it seems difficult that tuberculosis could have expanded and been maintained during the early stages of human expansion, as groups of migrants were probably of a too small size to enable the co-existence of the pathogen and its host (particularly in the conditions that must have accompanied the first steps of the settlement in a new area). To counter this objection, some may say that the characteristics of the tuberculosis bacillus when it appeared (particularly in terms of virulence) were most probably different from the characteristics of today's strains, which could allow endemicity even in small-sized populations. However, "*M. canettii*", despite being still in the environment and not transmissible from human to human, has a virulence that is close to that of *M. tuberculosis* when it infects people. This tends

to invalidate the hypothesis of a lesser virulence of *M. tuberculosis* during the early stages of its emergence. In contrast, according to "*M. canettii*" behavior, one might expect that the earlier *M. tuberculosis* strains were less able to spread from human to human, *i.e.* be less contagious rather than less virulent. As the infection by a smooth strain causes a disease similar to the one caused by a member of the MTBC, this proves that the characteristics of such an infection are not intrinsic to the obligatory pathogen character of *M. tuberculosis,* and thus have not necessarily been acquired after a long co-evolution between the bacterium and its host. In other words, the latent infection by *M. tuberculosis* would not be a characteristic specific to that pathogen but a "side effect", linked to the interaction between the infectious behavior of the bacillus and the defense mechanisms of the human body. Consequently it seems difficult to convincingly support the hypothesis of a very ancient origin of the *M. tuberculosis* complex on the simple basis of characteristics that could be explained by an expansion taking place while the population density around the world was still low, as for instance the propensity of the disease to stay at a latent stage for many years in infected individuals. Finally and as previously discussed, the parallel established by Comas *et al.* between the structure of the mitochondrial haplogroups tree and the *M. tuberculosis* complex phylogenetic tree, which is used as a major argument to anchor temporally the proposed model, is not convincing. The mitochondrial tree is nested, in a way which does not fit with the MTBC tree.

If the validity of the hypothesis proposed by Comas *et al.* seems questionable, it is then necessary to consider the elements supporting the dating of a new model of emergence, based on the link between "*M. canettii*" and the MTBC, as well as on the importance of the location of lineage 7. According to the points presented previously, North-East Ethiopia is the most likely cradle place for the complex, and lineage 7 would be the contemporary representative that has stayed in this area while the other lineages have diversified during their expansion into the whole world. Based on the evolution of the density of human populations, and in relation with the studies cited previously about the size of population required for the endemicity of tuberculosis, it also seems likely that tuberculosis has spread in conjunction with the Neolithic

Demographic Transition (NDT) that started approximately 10,000 years ago. It is believed that in the Sahara region, as well as in the East of Africa, the NDT would have started around 7,000 BCE. What characterizes the NDT is that it extended over a long period of time (several centuries), while different socio-cultural modifications took place, leading to a fully settled agricultural civilization characterized by a certain number of technological elements (for instance the use of ceramics) (Bocquet-Appel, 2011). The first villages probably gathered a few hundreds of individuals, which created the conditions for maintaining a tuberculosis epidemic. The establishment of settlements led to the use of fixed water sources, enabling the appearance of a disease focus, which could then spread to the other members of the community. These water sources shared by the whole community (in contrast with flowing body of water mostly used by hunter-gatherers) could have facilitated the transit of the ancestor of the *M. tuberculosis* complex from an aquatic reservoir to human hosts in accordance with the characteristics of "*M. canettii*" and the other close environmental mycobacteria (*M. marinum*, *M. ulcerans*). It is reasonable to assume that, similarly to what is observed nowadays with the smooth strains, the ancestor of the MTBC first caused "punctual" infections that were not transmissible between humans, in those first sedentary settlements. The acquisition of transmissibility could mark the speciation of the complex, through the isolation of this new bacterial ecotype vis-à-vis the environment.

This leads to the question of the time frame, in the hypothesis that these events took place in Ethiopia, and of the elements explaining the gradation in the departures, observed when the splits are projected on the linear tree that links "*M. canettii*" to lineage 7. As mentioned previously, it is estimated that in this part of the world the NDT has started around 7,000 years BCE, which, according to our model and extrapolating from short-term substitution rates, would be the age of the MRCA of the MTBC. Then it is necessary to look at what is known about the history of this region of Africa, in order to see if the expansion of the different lineages forming the complex seems possible. In terms of civilization, it is believed that Ethiopia was the center of a kingdom called Land of Punt (in Egyptian sources). Few direct evidence from this

civilization are remaining, but there are numerous references to it in archeological remains from the Egyptian civilization. The Land of Punt was the purveyor of several luxury items for Egypt, namely gold, ivory, as well as skins from felines or myrrh (Pankhurst, 1997). Based on the first appearance in Egypt of this latest product, it is likely that the first contacts between Punt and Egypt took place around 3,500 BCE. At this time the trade was probably made by land (Figure 23).

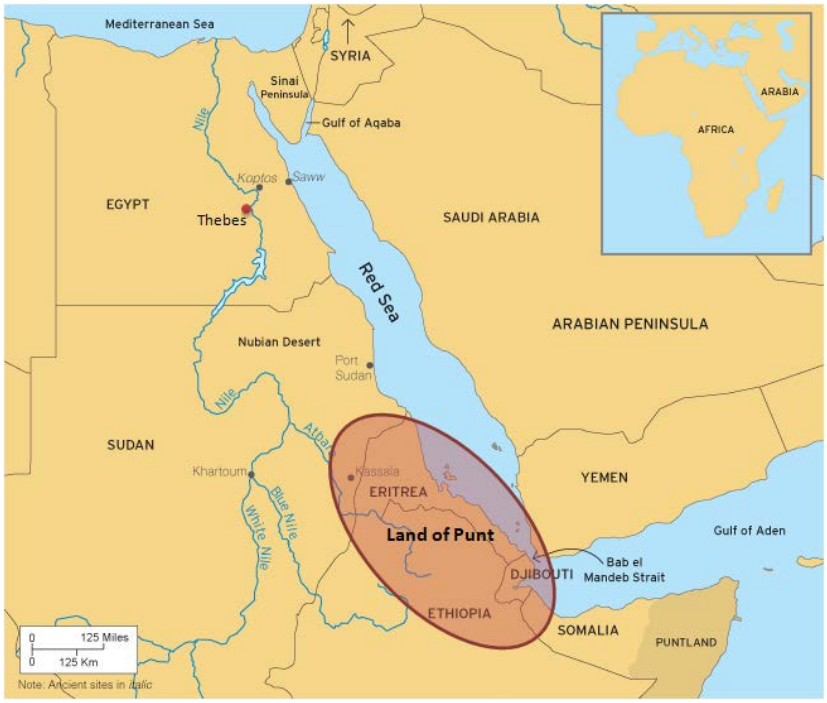


Figure 23: Probable location of the ancient Land of Punt.

The red area indicates this putative location, based on archeological evidence from Egyptian sources relating the trade with this kingdom.

However due to the harshness of the travel, be it by land or by sea, it appears that during the following millennia the contacts have been pursued more sporadically. This is one element that could fit with the transition identified by Zink *et al.* from superlineage 1 (lineages 5/6) to superlineage 3 (lineages 2, 3 & 4) in their analysis of infections caused by *M. tuberculosis* in Egyptian mummies. It seems also that this trade ended abruptly at some point, due to major cultural changes in the Ethiopian region, the Land of Punt being replaced by the Aksumite kingdom. Lineage 7 would have been preserved in the region of origin of the MTBC, because the populations of the Ethiopian highlands have lived in self-sufficiency, probably since the cultural changes that have accompanied the end of the Land of Punt. This isolation and this peculiarity

of the highlands region in Ethiopia has been detailed by the botanist Harlan (Harlan, 1969), who notes that a number of ancient agricultural practices, as well as many plants that correspond to ancestral forms of contemporary cultivated crops, are still present in the region.

Such a proposal for dating the MRCA of the MTBC can seem incompatible with the conclusions of some ancient DNA studies. For instance, Hershkovitz and col. (Hershkovitz *et al.*, 2008) identified *M. tuberculosis* DNA in the Eastern Mediterranean around 7,000 years BCE, i.e. at a time comparable with the proposed emergence of the MTBC in the East of Africa. The two described strains, coming from a mother and her newborn child, seem to present a spoligotype possessing most of the spacers, as well as a deleted TbD1 region, characteristic of (modern) superlineage 3. The DNA analysis is, however, subject to caution because of the age of the samples. For instance, the authors indicate that the spoligotyping amplifications (requiring the amplification of approx 70 bp fragments) were difficult to reproduce, whereas the TbD1 PCR was successful, producing a strong 128 bp amplification product, which is paradoxical.

Several points still need to be addressed in order to obtain new arguments in favor of an ancient origin scenario (70 kY) or a more recent one. A major argument could come from the sequencing of a significant proportion of an ancient strain, for instance one from an Egyptian mummy or the American bison. The projection of such a strain onto the tree of the complex obtained from a robust set of SNPs could enable to really calibrate the comparisons linking time and branch length, and thus to validate one model or the other depending on the position of the projection.

While waiting for such results, complementary approaches can be envisioned. The key question is the validity of the extrapolation of short term mutation rate of *M. tuberculosis* to longer time scales (Morelli *et al.*, 2010; Ford *et al.*, 2011; Comas *et al.*, 2013). A way to try to answer this problem is to take advantage of the availability of extensive MTBC genome sequence data to examine in detail the evolution of the dN/dS ratio between the different segments of the MTBC tree (Figure 22). Indeed, if in the long run the applied selective pressures change, thereby influencing the annual substitution rate, this change should be reflected on the dN/dS ratio for the branches

of the tree, depending whether they are terminal or more inside of the tree. If a global tendency exists that justifies a variation in the mutation rate, then the mutation rate inferred from outbreak events cannot be extended for the global evolution of the complex. On the contrary, in the absence of a tendency in the evolution of the dN/dS ratio, it is possible to suppose, as proposed by Ford *et al.*, that the mutation rate in *M. tuberculosis* is nearly always the same, thus allowing to evaluate the age of the MRCA of the complex by extrapolating from the observed branch length (and by using an outbreak as a calibration tool), as discussed above.

5) Conclusion: about the use of SNPs for phylogenetic purposes and the perspectives of high-throughput sequencing.

As observed from the study of the MTBC, the use of selected SNPs from whole genome comparisons is a very good phylogenetic tool because of its high resolutive power and low homoplasy. In practice, almost each strain, even within outbreaks, is individualized and possesses its own set of characteristic SNPs. This high resolution is particularly advantageous for the study of strains from epidemic events (such as "*M. canettii*" clone A). Such strains are usually clustered in a single genotype by first-line genotyping methods (for instance spoligotyping or MLVA). This is also the case for the study of evolutionary phenomena (such as the emergence of the MTBC) because of the mechanisms that govern the emergence and the selection of point mutations.

This new methodology has already replaced previous methods for several reasons. First of all the cost of draft whole genome sequencing of large number of genomes has dramatically decreased. Then the processing of the datasets is quickly becoming much easier, owing to the development of automated pipelines and simpler software. These developments take advantage of pilot research studies showing how to select SNPs appropriate for phylogenetic investigations. This opens the way to the making of either large databases that could enable to share SNPs data, or alternatively of software analyzing on-the-fly raw genomic sequence data available in depositories.

SNPs are particularly adapted to elucidate the phylogenetic relationships between closely related species, as illustrated with *M. tuberculosis* and "*M. canettii*". Several other human pathogens show the same type of population structure, with a close link with an environmental species. It is the case for instance for two of the most dangerous pathogens namely *Bacillus anthracis* (recently emerged from *Bacillus cereus*) and *Yersinia pestis* (recently emerged from *Yersinia pseudotuberculosis*). As for the MTBC, it is important to try and understand the evolutionary history of these pathogens, in order to identify the characteristics that made them such threats for humans. It could indeed open the way for new means to fight against them, as well as to control the emergence of new bacteria that have not yet acquired this level of

pathogenicity. The methods described in this work will have to be adapted to each case. Indeed the clonality, which is a characteristic of the MTBC members *stricto sensu* (i.e. not including "*M. canettii*"), may not apply to other species, and the evolutionary speed (i.e. the branch length) in the different lineages may not necessarily be the same, which could impact the analysis techniques, as well as the conclusions deduced from the study of the SNPs.

When time and money is not a major issue (i.e. for pilot studies in a research context), the study of SNPs already complements the traditional typing techniques, which are only used in a pre-screening phase as illustrated here. Sequencing technologies continue to improve steadily, in such a way that in a near future, it may be possible to obtain a "true" whole genome directly after sequencing for the cost of today's draft genomes. This new major change has been initiated by the development of sequencers that produce long reads, at a high-throughput (as opposed to short reads produced by the current technologies such as Illumina). One of the leading company in this field is Pacific Biosciences, which offers a cost around $3,000 for the long-reads sequencing, thus opening the way to almost complete bacterial genomes. Two other companies are currently developing promising technologies to access truly complete genomes. There is first Moleculo (a society that has been bought by Illumina) that developed a process to artificially reconstruct long reads from sequencing data produced with the Illumina technology. This approach is a way to get rid of the problem of interspersed repetitive regions along the genome. However it encounters still some troubles with tandemly repeated regions of a large size (as for instance some VNTRs). But even if the prices have not yet been clearly set, this method might be able to challenge Pacific Biosciences. The second alternative comes from single molecule sequencing through nanopore as performed by the sequencers from Oxford Nanopore. These machines are just entering the field, and comparative studies will enable to compare their performances with the other existing technologies. In all cases it seems that the transition to long reads is happening now, and that the standard will soon be the complete reconstruction of studied genomes. One side aspect of this development is that classical typing information, such as VNTR size, will be more easily accessible

than with the current technologies. This will be of great use to insure the compatibility with anterior typing data, and more importantly to maintain elementary genotyping technologies such as MLVA, which can be applied for a few Euros and no specific equipment to crude DNA preparations.

This new technological step forward induces several other remarks. Firstly, the study of degraded samples (because of transport or extraction conditions or because it is ancient DNA) will still lead to the sequencing of small fragments that will need to be assembled. This will always be of importance when such fragmentary genomes will be compared with entire ones (just as with today's sequencing). Consequently the development of powerful assembly algorithms will remain useful for these particular applications, in order to be able to compare genomes coming from different sources. It is nevertheless possible to imagine ways to use the data gathered from this type of analysis through other means, for instance by trying to align the fragmentary DNA on targets from a "knowledge database" that would regroup the loci under investigation, or by aligning the DNA on a putative ancestral genome reconstructed *in silico*. Nonetheless, for strains that can be cultivated, it seems obvious that complete sequencing using some kind of long reads technology will become the next gold-standard.

If this solves partly the problem of the assembly of genomes, the remaining difficulties will be found at the level of the informatics treatment of the produced sequences (i.e. to generate comparisons). Indeed even when taking into account the suppression of the assembly step, the alignment of a large number of genomes in order to compare them is a limiting step. This is without mentioning the fact that the technical difficulties for alignment (mentioned in the introduction) will remain and could cause the loss of a part of the advantages gained from obtaining complete genome sequences (because of the alignment of segments).

Besides, today, even if the production of genomic data represents a large part in biological research, not so many biologists have the basic knowledge in informatics required to be able to process them directly. In addition, because of the ever-increasing amount of this kind of data (mainly because of the lowering of the

sequencing costs), there is probably not enough infrastructures to externalize sequence analysis. It is therefore necessary to develop new analysis procedures and adapted tools that could permit both an important automation (in order to save time), and an ease of use (to make it accessible to non-specialists). Since the earliest developments of this work, some changes have taken place in terms of computer treatment of NGS data (and genomic data in general). However most of them are fairly minor, as they consist essentially in the improvement of previously existing tools or algorithms. For instance, for *de novo* assemblies, most algorithms have been modified to take into account the augmentation of the reads length produced by the Illumina technology. There are not yet many tools that enable easy manipulations of datasets, and many high-level applications still require extensive programming knowledge to be performed in an effective way, mostly in terms of the time-cost implied to perform an analysis.

IV - Bibliography

Abadia, E., J. Zhang, V. Ritacco, K. Kremer, R. Ruimy, L. Rigouts, H. M. Gomes, A. R. Elias, M. Fauville-Dufaux, K. Stoffels, V. Rasolofo-Razanamparany, D. Garcia de Viedma, M. Herranz, S. Al-Hajoj, N. Rastogi, C. Garzelli, E. Tortoli, P. N. Suffys, D. van Soolingen, G. Refregier, et al. (2011). "The use of microbead-based spoligotyping for *Mycobacterium tuberculosis* complex to evaluate the quality of the conventional method: providing guidelines for Quality Assurance when working on membranes." BMC Infect Dis **11**: 110.

Avery, O. T., C. M. Macleod and M. McCarty (1944). "Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from *Pneumococcus* Type Iii." J Exp Med **79**(2): 137-158.

Balasubramanian, V., E. H. Wiegeshaus and D. W. Smith (1994). "Mycobacterial infection in guinea pigs." Immunobiology **191**(4-5): 395-401.

Barrera, L. (2007). The Basics of Clinical Bacteriology. Tuberculosis 2007, From basic science to patient care.

Becq, J., M. C. Gutierrez, V. Rosas-Magallanes, J. Rauzier, B. Gicquel, O. Neyrolles and P. Deschavanne (2007). "Contribution of horizontally acquired genomic islands to the evolution of the tubercle bacilli." Mol Biol Evol **24**(8): 1861-1871.

Behr, M. A. (2013). "Evolution of *Mycobacterium tuberculosis*." Adv Exp Med Biol **783**: 81-91.

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, et al. (2008). "Accurate whole human genome sequencing using reversible terminator chemistry." Nature **456**(7218): 53-59.

Blouin, Y., G. Cazajous, C. Dehan, C. Soler, R. Vong, M. O. Hassan, Y. Hauck, C. Boulais, D. Andriamanantena, C. Martinaud, E. Martin, C. Pourcel and G. Vergnaud (2014). "Progenitor "*Mycobacterium canettii*" clone responsible for lymph node tuberculosis epidemic, Djibouti." Emerg Infect Dis **20**(1): 21-28.

Blouin, Y., Y. Hauck, C. Soler, M. Fabre, R. Vong, C. Dehan, G. Cazajous, P. L. Massoure, P. Kraemer, A. Jenkins, E. Garnotel, C. Pourcel and G. Vergnaud (2012). "Significance of the identification in the Horn of Africa of an exceptionally deep branching Mycobacterium tuberculosis clade." PLoS One **7**(12): e52841.

Bocquet-Appel, J. P. (2011). "When the world's population took off: the springboard of the Neolithic Demographic Transition." Science **333**(6042): 560-561.

Boisvert, S., F. Laviolette and J. Corbeil (2010). "Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies." J Comput Biol **17**(11): 1519-1533.

Bos, K. I., V. J. Schuenemann, G. B. Golding, H. A. Burbano, N. Waglechner, B. K. Coombes, J. B. McPhee, S. N. DeWitte, M. Meyer, S. Schmedes, J. Wood, D. J. Earn, D. A. Herring, P. Bauer, H. N. Poinar and J. Krause (2011). "A draft genome of *Yersinia pestis* from victims of the Black Death." Nature **478**(7370): 506-510.

Brosch, R., S. V. Gordon, M. Marmiesse, P. Brodin, C. Buchrieser, K. Eiglmeier, T. Garnier, C. Gutierrez, G. Hewinson, K. Kremer, L. M. Parsons, A. S. Pym, S. Samper, D. van Soolingen and S. T. Cole (2002). "A new evolutionary scenario for the *Mycobacterium tuberculosis* complex." Proc Natl Acad Sci U S A **99**(6): 3684-3689.

Brudey, K., J. R. Driscoll, L. Rigouts, W. M. Prodinger, A. Gori, S. A. Al-Hajoj, C. Allix, L. Aristimuno, J. Arora, V. Baumanis, L. Binder, P. Cafrune, A. Cataldi, S. Cheong, R. Diel, C. Ellermeier, J. T. Evans, M. Fauville-Dufaux, S. Ferdinand, D. Garcia de Viedma, et al. (2006). "*Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology." BMC Microbiol **6**: 23.

Burns, J., Burrow, S., Genovese, E., Pumphery, M., Sims, E., Thomson, N.J. (2003). Other communicable diseases. The health of indigenous Australians, Oxford University Press**:** 397-441.

Camus, J. C., M. J. Pryor, C. Medigue and S. T. Cole (2002). "Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv." Microbiology **148**(Pt 10): 2967-2973.

Cardoso Leão, S. and F. Portaels (2007). History. Tuberculosis 2007, From basic science to patient care.

Chalke, H. D. (1959). "Some historical aspects of tuberculosis." Public Health **74**: 83-95.

Chevreux, B. (2005). MIRA: An Automated Genome and EST Assembler. Department of Molecular Biophysics. Heidelberg, The Medical Faculty of Heidelberg. **Doctor scientiarum humanarum (Dr.sc.hum.):** 171.

Cohan, F. M. (2001). "Bacterial species and speciation." Syst Biol **50**(4): 513-524.

Cole, S. T. and B. G. Barrell (1998). "Analysis of the genome of *Mycobacterium tuberculosis* H37Rv." Novartis Found Symp **217**: 160-172; discussion 172-167.

Comas, I., J. Chakravartti, P. M. Small, J. Galagan, S. Niemann, K. Kremer, J. D. Ernst and S. Gagneux (2010). "Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved." Nat Genet **42**(6): 498-503.

Comas, I., M. Coscolla, T. Luo, S. Borrell, K. E. Holt, M. Kato-Maeda, J. Parkhill, B. Malla, S. Berg, G. Thwaites, D. Yeboah-Manu, G. Bothamley, J. Mei, L. Wei, S. Bentley, S. R. Harris, S. Niemann, R. Diel, A. Aseffa, Q. Gao, et al. (2013). "Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans." Nat Genet **45**(10): 1176-1182.

Covert, A. W., 3rd, R. E. Lenski, C. O. Wilke and C. Ofria (2013). "Experiments on the role of deleterious mutations as stepping stones in adaptive evolution." Proc Natl Acad Sci U S A **110**(34): E3171-3178.

Cowan, L. S., L. Diem, M. C. Brake and J. T. Crawford (2004). "Transfer of a *Mycobacterium tuberculosis* genotyping method, Spoligotyping, from a reverse line-blot hybridization, membrane-based assay to the Luminex multianalyte profiling system." J Clin Microbiol **42**(1): 474-477.

Crick, F. H. (1954). "The Complementary Structure of DNA." Proc Natl Acad Sci U S A **40**(8): 756-758.

Croucher, N. J., S. R. Harris, C. Fraser, M. A. Quail, J. Burton, M. van der Linden, L. McGee, A. von Gottberg, J. H. Song, K. S. Ko, B. Pichon, S. Baker, C. M. Parry, L. M. Lambertsen, D. Shahinas, D. R. Pillai, T. J. Mitchell, G. Dougan, A. Tomasz, K. P. Klugman, et al. (2011). "Rapid pneumococcal evolution in response to clinical interventions." Science **331**(6016): 430-434.

Daffé, M., C. Lacave, M. A. Laneelle and G. Laneelle (1987). "Structure of the major triglycosyl phenol-phthiocerol of *Mycobacterium tuberculosis* (strain Canetti)." Eur J Biochem **167**(1): 155-160.

Daniel, T. M. (2006). "The history of tuberculosis." Respir Med **100**(11): 1862-1870.

Darling, A. E., B. Mau and N. T. Perna (2010). "progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement." PLoS One **5**(6): e11147.

Darwin, C. (1859). On the Origin of Species.

Dean, G. S., S. G. Rhodes, M. Coad, A. O. Whelan, P. J. Cockle, D. J. Clifford, R. G. Hewinson and H. M. Vordermeier (2005). "Minimum infective dose of *Mycobacterium bovis* in cattle." Infect Immun **73**(10): 6467-6471.

Demay, C., B. Liens, T. Burguiere, V. Hill, D. Couvin, J. Millet, I. Mokrousov, C. Sola, T. Zozio and N. Rastogi (2012). "SITVITWEB--a publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology." Infect Genet Evol **12**(4): 755-766.

Deshayes, C., E. Perrodou, D. Euphrasie, E. Frapy, O. Poch, P. Bifani, O. Lecompte and J. M. Reyrat (2008). "Detecting the molecular scars of evolution in the *Mycobacterium tuberculosis* complex by analyzing interrupted coding sequences." BMC Evol Biol **8**: 78.

Diamond, J. (1999). Guns, Germs and Steel. New York.

Elvin, M. and C. Liu (1998). Sediments of Time: Environment and Society in Chinese History.

Engel, H. W., L. G. Berwald and A. H. Havelaar (1980). "The occurrence of *Mycobacterium kansasii* in tapwater." Tubercle **61**(1): 21-26.

Fabre, M., Y. Hauck, C. Soler, J. L. Koeck, J. van Ingen, D. van Soolingen, G. Vergnaud and C. Pourcel (2010). "Molecular characteristics of "*Mycobacterium canettii*" the smooth *Mycobacterium tuberculosis* bacilli." Infect Genet Evol **10**(8): 1165-1173.

Fabre, M., J. L. Koeck, P. Le Flèche, F. Simon, V. Hervé, G. Vergnaud and C. Pourcel (2004). "High genetic diversity revealed by variable-number tandem repeat genotyping and analysis of hsp65 gene polymorphism in a large collection of "*Mycobacterium canettii*" strains indicates that the *M. tuberculosis* complex is a recently emerged clone of "*M. canettii*"." J Clin Microbiol **42**(7): 3248-3255.

Filliol, I., J. R. Driscoll, D. Van Soolingen, B. N. Kreiswirth, K. Kremer, G. Valetudie, D. D. Anh, R. Barlow, D. Banerjee, P. J. Bifani, K. Brudey, A. Cataldi, R. C. Cooksey, D. V. Cousins, J. W. Dale, O. A. Dellagostin, F. Drobniewski, G. Engelmann, S. Ferdinand, D. Gascoyne-Binzi, et al. (2002). "Global distribution of *Mycobacterium tuberculosis* spoligotypes." Emerg Infect Dis **8**(11): 1347-1349.

Filliol, I., J. R. Driscoll, D. van Soolingen, B. N. Kreiswirth, K. Kremer, G. Valetudie, D. A. Dang, R. Barlow, D. Banerjee, P. J. Bifani, K. Brudey, A. Cataldi, R. C. Cooksey, D. V. Cousins, J. W. Dale, O. A. Dellagostin, F. Drobniewski, G. Engelmann, S. Ferdinand, D. Gascoyne-Binzi, et al. (2003). "Snapshot of moving and expanding clones of *Mycobacterium tuberculosis* and their global distribution assessed by spoligotyping in an international study." J Clin Microbiol **41**(5): 1963-1970.

Filliol, I., A. S. Motiwala, M. Cavatore, W. Qi, M. H. Hazbon, M. Bobadilla del Valle, J. Fyfe, L. Garcia-Garcia, N. Rastogi, C. Sola, T. Zozio, M. I. Guerrero, C. I. Leon, J. Crabtree, S. Angiuoli, K. D. Eisenach, R. Durmaz, M. L. Joloba, A. Rendon, J. Sifuentes-Osornio, et al. (2006). "Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic

accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set." J Bacteriol **188**(2): 759-772.

Firdessa, R., S. Berg, E. Hailu, E. Schelling, B. Gumi, G. Erenso, E. Gadisa, T. Kiros, M. Habtamu, J. Hussein, J. Zinsstag, B. D. Robertson, G. Ameni, A. J. Lohan, B. Loftus, I. Comas, S. Gagneux, R. Tschopp, L. Yamuah, G. Hewinson, et al. (2013). "Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia." Emerg Infect Dis **19**(3): 460-463.

Fleagle, J. G., Z. Assefa, F. H. Brown and J. J. Shea (2008). "Paleoanthropology of the Kibish Formation, southern Ethiopia: Introduction." J Hum Evol **55**(3): 360-365.

Ford, C. B., P. L. Lin, M. R. Chase, R. R. Shah, O. Iartchouk, J. Galagan, N. Mohaideen, T. R. Ioerger, J. C. Sacchettini, M. Lipsitch, J. L. Flynn and S. M. Fortune (2011). "Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection." Nat Genet **43**(5): 482-486.

Frothingham, R. and W. A. Meeker-O'Connell (1998). "Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats." Microbiology **144 (Pt 5)**: 1189-1196.

Gagneux, S. and P. M. Small (2007). "Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development." Lancet Infect Dis **7**(5): 328-337.

Gardy, J. L., J. C. Johnston, S. J. Ho Sui, V. J. Cook, L. Shah, E. Brodkin, S. Rempel, R. Moore, Y. Zhao, R. Holt, R. Varhol, I. Birol, M. Lem, M. K. Sharma, K. Elwood, S. J. Jones, F. S. Brinkman, R. C. Brunham and P. Tang (2011). "Whole-genome sequencing and social-network analysis of a tuberculosis outbreak." N Engl J Med **364**(8): 730-739.

Gurven, M., H. Kaplan and A. Z. Supa (2007). "Mortality experience of Tsimane Amerindians of Bolivia: regional variation and temporal trends." Am J Hum Biol **19**(3): 376-398.

Gutierrez, M. C., S. Brisse, R. Brosch, M. Fabre, B. Omais, M. Marmiesse, P. Supply and V. Vincent (2005). "Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*." PLoS Pathog **1**(1): e5.

Hamilton, M. J., B. T. Milne, R. S. Walker, O. Burger and J. H. Brown (2007). "The complex structure of hunter-gatherer social networks." Proc Biol Sci **274**(1622): 2195-2202.

Harlan, J. (1969). Ethiopia: A Center of Diversity. Annual meeting of the Society for Economic Botany. Columbus, Ohio.

Harshey, R. M. and T. Ramakrishnan (1977). "Rate of ribonucleic acid chain growth in *Mycobacterium tuberculosis* H37Rv." J Bacteriol **129**(2): 616-622.

Hernández-Pando, R., R. Chacón-Salinas, J. Serafín-López and I. Estrada (2007). Immunology, Pathogenesis, Virulence. Tuberculosis 2007, From basic science to patient care.

Hershberg, R., M. Lipatov, P. M. Small, H. Sheffer, S. Niemann, S. Homolka, J. C. Roach, K. Kremer, D. A. Petrov, M. W. Feldman and S. Gagneux (2008). "High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography." PLoS Biol **6**(12): e311.

Hershey, A. D. and M. Chase (1952). "Independent functions of viral protein and nucleic acid in growth of bacteriophage." J Gen Physiol **36**(1): 39-56.

Hershkovitz, I., H. D. Donoghue, D. E. Minnikin, G. S. Besra, O. Y. Lee, A. M. Gernaey, E. Galili, V. Eshed, C. L. Greenblatt, E. Lemma, G. K. Bar-Gal and M. Spigelman (2008). "Detection and molecular characterization of 9,000-year-old *Mycobacterium tuberculosis* from a Neolithic settlement in the Eastern Mediterranean." PLoS One **3**(10): e3426.

Hett, E. C. and E. J. Rubin (2008). "Bacterial growth and cell division: a mycobacterial perspective." Microbiol Mol Biol Rev **72**(1): 126-156, table of contents.

Jenner, E. (1798). An Inquiry into the Causes and Effects of the Variolæ Vaccinæ. London.

Johnson, P. D., J. B. Carlin, C. M. Bennett, P. D. Phelan, M. Starr, J. Hulls and T. M. Nolan (1998). "Prevalence of tuberculosis infection in Melbourne secondary school students." Med J Aust **168**(3): 106-110.

Kamerbeek, J., L. Schouls, A. Kolk, M. van Agterveld, D. van Soolingen, S. Kuijper, A. Bunschoten, H. Molhuizen, R. Shaw, M. Goyal and J. van Embden (1997). "Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology." J Clin Microbiol **35**(4): 907-914.

Koch, R. (1882). "Die Aetiologie der Tuberkulose." Berliner Klinische Wochenschrift **19**(15): 221-230.

Koeck, J. L., M. Fabre, F. Simon, M. Daffe, E. Garnotel, A. B. Matan, P. Gerome, J. J. Bernatas, Y. Buisson and C. Pourcel (2011). "Clinical characteristics of the smooth tubercle bacilli '*Mycobacterium canettii*' infection suggest the existence of an environmental reservoir." Clin Microbiol Infect **17**(7): 1013-1019.

Kritski, A. and F. A. Fiuza de Melo (2007). Tuberculosis in Adults. Tuberculosis 2007, From basic science to patient care.

Kröpelin, S., D. Verschuren, A. M. Lezine, H. Eggermont, C. Cocquyt, P. Francus, J. P. Cazet, M. Fagot, B. Rumes, J. M. Russell, F. Darius, D. J. Conley, M. Schuster, H. von Suchodoletz and D. R. Engstrom (2008). "Climate-driven ecosystem succession in the Sahara: the past 6000 years." Science **320**(5877): 765-768.

Kryazhimskiy, S. and J. B. Plotkin (2008). "The population genetics of dN/dS." PLoS Genet **4**(12): e1000304.

Laennec, R. T. H. (1819). De l'Auscultation médiate. Paris.

Le Flèche, P., M. Fabre, F. Denoeud, J. L. Koeck and G. Vergnaud (2002). "High resolution, on-line identification of strains from the Mycobacterium tuberculosis complex based on tandem repeat typing." BMC Microbiol **2**: 37.

Lemassu, A., V. V. Levy-Frebault, M. A. Laneelle and M. Daffe (1992). "Lack of correlation between colony morphology and lipooligosaccharide content in the Mycobacterium tuberculosis complex." J Gen Microbiol **138**(7): 1535-1541.

Liu, H., F. Prugnolle, A. Manica and F. Balloux (2006). "A geographically explicit genetic model of worldwide human-settlement history." Am J Hum Genet **79**(2): 230-237.

Marshall, F. and E. Hildebrand (2002). "Cattle Before Crops: The Beginnings of Food Production in Africa." Journal of World Prehistory **16**(2).

Marsollier, L., R. Robert, J. Aubry, J. P. Saint Andre, H. Kouakou, P. Legras, A. L. Manceau, C. Mahaza and B. Carbonnelle (2002). "Aquatic insects as a vector for *Mycobacterium ulcerans*." Appl Environ Microbiol **68**(9): 4623-4628.

Martín, C., F. Bigi and B. Gicquel (2007). New Vaccines against Tuberculosis. Tuberculosis 2007, From basic science to patient care.

Maxam, A. M. and W. Gilbert (1977). "A new method for sequencing DNA." Proc Natl Acad Sci U S A **74**(2): 560-564.

Mazars, E., S. Lesjean, A. L. Banuls, M. Gilbert, V. Vincent, B. Gicquel, M. Tibayrenc, C. Locht and P. Supply (2001). "High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology." Proc Natl Acad Sci U S A **98**(4): 1901-1906.

Mba Medie, F., I. Ben Salah, B. Henrissat, D. Raoult and M. Drancourt (2011). "*Mycobacterium tuberculosis* complex mycobacteria as amoeba-resistant organisms." PLoS One **6**(6): e20499.

McGrath, J. (1988). "Social Networks of Disease Spread in the Lower Illinois Valley: A Simulation Approach." Am J Phys Anthropol **77**: 483-496.

Miltgen, J., M. Morillon, J. L. Koeck, A. Varnerot, J. F. Briant, G. Nguyen, D. Verrot, D. Bonnet and V. Vincent (2002). "Two cases of pulmonary tuberculosis caused by *Mycobacterium tuberculosis* subsp *canetti*." Emerg Infect Dis **8**(11): 1350-1352.

Morelli, G., X. Didelot, B. Kusecek, S. Schwarz, C. Bahlawane, D. Falush, S. Suerbaum and M. Achtman (2010). "Microevolution of *Helicobacter pylori* during prolonged infection of single hosts and within families." PLoS Genet **6**(7): e1001036.

Morelli, G., Y. Song, C. J. Mazzoni, M. Eppinger, P. Roumagnac, D. M. Wagner, M. Feldkamp, B. Kusecek, A. J. Vogler, Y. Li, Y. Cui, N. R. Thomson, T. Jombart, R. Leblois, P. Lichtner, L. Rahalison, J. M. Petersen, F. Balloux, P. Keim, T. Wirth, et al. (2010). "*Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity." Nat Genet **42**(12): 1140-1143.

Nicklisch, N., F. Maixner, R. Ganslmeier, S. Friederich, V. Dresely, H. Meller, A. Zink and K. W. Alt (2012). "Rib lesions in skeletons from early neolithic sites in Central Germany: on the trail of tuberculosis at the onset of agriculture." Am J Phys Anthropol **149**(3): 391-404.

Nyrén, P. (2007). "The history of pyrosequencing." Methods Mol Biol **373**: 1-14.

Pankhurst, R. (1997). The Ethiopian Borderlands.

Pepperell, C. S., A. M. Casto, A. Kitchen, J. M. Granka, O. E. Cornejo, E. C. Holmes, B. Birren, J. Galagan and M. W. Feldman (2013). "The role of selection in shaping diversity of natural M. tuberculosis populations." PLoS Pathog **9**(8): e1003543.

Perego, U. A., A. Achilli, N. Angerhofer, M. Accetturo, M. Pala, A. Olivieri, B. Hooshiar Kashani, K. H. Ritchie, R. Scozzari, Q. P. Kong, N. M. Myres, A. Salas, O. Semino, H. J. Bandelt, S. R. Woodward and A. Torroni (2009). "Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups." Curr Biol **19**(1): 1-8.

Perrodou, E., C. Deshayes, J. Muller, C. Schaeffer, A. Van Dorsselaer, R. Ripp, O. Poch, J. M. Reyrat and O. Lecompte (2006). "ICDS database: interrupted CoDing sequences in prokaryotic genomes." Nucleic Acids Res **34**(Database issue): D338-343.

Pfyffer, G. E., R. Auckenthaler, J. D. van Embden and D. van Soolingen (1998). "*Mycobacterium canettii*, the smooth variant of *M. tuberculosis*, isolated from a Swiss patient exposed in Africa." Emerg Infect Dis **4**(4): 631-634.

Portaels, F., P. Elsen, A. Guimaraes-Peres, P. A. Fonteyne and W. M. Meyers (1999). "Insects in the transmission of *Mycobacterium ulcerans* infection." Lancet **353**(9157): 986.

Pourcel, C., G. Salvignol and G. Vergnaud (2005). "CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies." Microbiology **151**(Pt 3): 653-663.

Ronaghi, M., S. Karamohamed, B. Pettersson, M. Uhlen and P. Nyren (1996). "Real-time DNA sequencing using detection of pyrophosphate release." Anal Biochem **242**(1): 84-89.

Rothschild, B. M., L. D. Martin, G. Lev, H. Bercovier, G. K. Bar-Gal, C. Greenblatt, H. Donoghue, M. Spigelman and D. Brittain (2001). "*Mycobacterium tuberculosis* complex DNA from an extinct bison dated 17,000 years before the present." Clin Infect Dis **33**(3): 305-311.

Salah, I. B., E. Ghigo and M. Drancourt (2009). "Free-living amoebae, a training field for macrophage resistance of mycobacteria." Clin Microbiol Infect **15**(10): 894-905.

Sanger, F., S. Nicklen and A. R. Coulson (1977). "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci U S A **74**(12): 5463-5467.

Smith, J. M., N. H. Smith, M. O'Rourke and B. G. Spratt (1993). "How clonal are bacteria?" Proc Natl Acad Sci U S A **90**(10): 4384-4388.

Smith, N. H. (2006). "A re-evaluation of *M. prototuberculosis*." PLoS Pathog **2**(9): e98.

Smith, N. H., R. G. Hewinson, K. Kremer, R. Brosch and S. V. Gordon (2009). "Myths and misconceptions: the origin and evolution of *Mycobacterium tuberculosis*." Nat Rev Microbiol **7**(7): 537-544.

Smith, N. H., K. Kremer, J. Inwald, J. Dale, J. R. Driscoll, S. V. Gordon, D. van Soolingen, R. G. Hewinson and J. M. Smith (2006). "Ecotypes of the *Mycobacterium tuberculosis* complex." J Theor Biol **239**(2): 220-225.

Sreevatsan, S., X. Pan, K. E. Stockbauer, N. D. Connell, B. N. Kreiswirth, T. S. Whittam and J. M. Musser (1997). "Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination." Proc Natl Acad Sci U S A **94**(18): 9869-9874.

Stamm, L. M. and E. J. Brown (2004). "*Mycobacterium marinum*: the generalization and specialization of a pathogenic mycobacterium." Microbes Infect **6**(15): 1418-1428.

Stinear, T. P., G. A. Jenkin, P. D. Johnson and J. K. Davies (2000). "Comparative genetic analysis of *Mycobacterium ulcerans* and *Mycobacterium marinum* reveals evidence of recent divergence." J Bacteriol **182**(22): 6322-6330.

Stinear, T. P., T. Seemann, P. F. Harrison, G. A. Jenkin, J. K. Davies, P. D. Johnson, Z. Abdellah, C. Arrowsmith, T. Chillingworth, C. Churcher, K. Clarke, A. Cronin, P. Davis, I. Goodhead, N. Holroyd, K. Jagels, A. Lord, S. Moule, K. Mungall, H. Norbertczak, et al. (2008). "Insights from the complete genome sequence of *Mycobacterium marinum* on the evolution of *Mycobacterium tuberculosis*." Genome Res **18**(5): 729-741.

Supply, P., M. Marceau, S. Mangenot, D. Roche, C. Rouanet, V. Khanna, L. Majlessi, A. Criscuolo, J. Tap, A. Pawlik, L. Fiette, M. Orgeur, M. Fabre, C. Parmentier, W. Frigui, R. Simeone, E. C. Boritsch, A. S. Debrie, E. Willery, D. Walker, et al. (2013). "Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*." Nat Genet **45**(2): 172-179.

Thierry, D., A. Brisson-Noel, V. Vincent-Levy-Frebault, S. Nguyen, J. L. Guesdon and B. Gicquel (1990). "Characterization of a *Mycobacterium tuberculosis* insertion sequence, IS6110, and its application in diagnosis." J Clin Microbiol **28**(12): 2668-2673.

Turesson, G. (1922). "The Species and the Variety as Ecological Units." Hereditas **3**(1): 100-113.

Turrill, W. B. (1946). "The Ecotype Concept." New Phytologist **45**(1): 34-43.

van Embden, J. D., M. D. Cave, J. T. Crawford, J. W. Dale, K. D. Eisenach, B. Gicquel, P. Hermans, C. Martin, R. McAdam, T. M. Shinnick and et al. (1993). "Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology." J Clin Microbiol **31**(2): 406-409.

van Embden, J. D., T. van Gorkom, K. Kremer, R. Jansen, B. A. van Der Zeijst and L. M. Schouls (2000). "Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria." J Bacteriol **182**(9): 2393-2401.

van Soolingen, D., P. E. de Haas, J. Haagsma, T. Eger, P. W. Hermans, V. Ritacco, A. Alito and J. D. van Embden (1994). "Use of various genetic markers in differentiation of

Mycobacterium bovis strains from animals and humans and for studying epidemiology of bovine tuberculosis." J Clin Microbiol **32**(10): 2425-2433.

van Soolingen, D., T. Hoogenboezem, P. E. de Haas, P. W. Hermans, M. A. Koedam, K. S. Teppema, P. J. Brennan, G. S. Besra, F. Portaels, J. Top, L. M. Schouls and J. D. van Embden (1997). "A novel pathogenic taxon of the *Mycobacterium tuberculosis* complex, Canetti: characterization of an exceptional isolate from Africa." Int J Syst Bacteriol **47**(4): 1236-1245.

Veyrier, F., D. Pletzer, C. Turenne and M. A. Behr (2009). "Phylogenetic detection of horizontal gene transfer during the step-wise genesis of *Mycobacterium tuberculosis*." BMC Evol Biol **9**: 196.

Villemin, J. A. (1865). "Cause et nature de la tuberculose: son inoculation de l'homme au lapin." Compt Rend Acad Sci **61**: 1012-1015.

Wagner, D. M., J. Klunk, M. Harbeck, A. Devault, N. Waglechner, J. W. Sahl, J. Enk, D. N. Birdsell, M. Kuch, C. Lumibao, D. Poinar, T. Pearson, M. Fourment, B. Golding, J. M. Riehm, D. J. Earn, S. Dewitte, J. M. Rouillard, G. Grupe, I. Wiechmann, et al. (2014). "*Yersinia pestis* and the Plague of Justinian 541-543 AD: a genomic analysis." Lancet Infect Dis.

Walker, T. M., C. L. Ip, R. H. Harrell, J. T. Evans, G. Kapatai, M. J. Dedicoat, D. W. Eyre, D. J. Wilson, P. M. Hawkey, D. W. Crook, J. Parkhill, D. Harris, A. S. Walker, R. Bowden, P. Monk, E. G. Smith and T. E. Peto (2013). "Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study." Lancet Infect Dis **13**(2): 137-146.

WHO (2013). Global tuberculosis report 2013.

Wirth, T., F. Hildebrand, C. Allix-Beguec, F. Wolbeling, T. Kubica, K. Kremer, D. van Soolingen, S. Rusch-Gerdes, C. Locht, S. Brisse, A. Meyer, P. Supply and S. Niemann (2008). "Origin, spread and demography of the *Mycobacterium tuberculosis* complex." PLoS Pathog **4**(9): e1000160.

Zerbino, D. R. and E. Birney (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome Res **18**(5): 821-829.

Zink, A. R., C. Sola, U. Reischl, W. Grabner, N. Rastogi, H. Wolf and A. G. Nerlich (2003). "Characterization of *Mycobacterium tuberculosis* complex DNAs from Egyptian mummies by spoligotyping." J Clin Microbiol **41**(1): 359-367.

Zysk, K. G. (1998). Medicine in the Veda.

V - Annexes

A) Methods:

Assemblies

For most analyses performed in this work (and for all SNPs determination), the type of assembly that was selected was "homology", for two reasons. The first one is that SNPs have to be selected from the core genome (i.e. shared by all strains being investigated) in order to be used as a phylogenetic marker, and the second one is linked with the studied organism. As all *M. tuberculosis* strains are closely related (i.e. genetically homogenous), this type of assembly was particularly well suited to their study. It permits also to identify regions absent in the strain studied compared to the reference strain, on which the assembly is performed. In an "homology" assembly, sequencing reads are mapped on a reference genome. The user is asked to set a number of parameters which will define criteria for mapping a read (average similarity, sequence quality), and criteria for calling an SNP (coverage).

For all MTBC analyses, strain H37Rv was used as reference strain, because its sequence had been the subject of a continuous correction process since its initial release (Cole *et al.*, 1998; Camus *et al.*, 2002), to make it the genomic gold-standard in the field of *M. tuberculosis* research. This is more comfortable when determining SNPs, as sequencing errors in the reference genome will lead to the inclusion of false variations and an excessive branch length towards the reference. The execution of an homology assembly with a software such as BioNumerics (BN, Applied-Maths, Belgium) is relatively straightforward. Using the Power Assembler module present in this version of the software (v 7), it is possible to construct an assembly pipeline from basic blocks that perform standard actions, like loading a file. An homology assembly is composed by a minimal skeleton of four steps: loading the reference sequence from the BN database, importing the reads from a file (for instance a fastq format file), assembling the reads on the reference sequence, and finally exporting the assembled target as a new sequence to be stored in the database. Some parameters need to be

empirically tested: minimum coverage required to validate a SNP, degree of consensus, sequence read similarity, read quality. As a rule, the better parameters will be those providing the lowest number of SNPs per kilobase of genome covered. No specific scripts have been developed for this part. The output of a mapping step is a sequence stored in the BN database of exactly the same size as the reference genome. Ns will have been inserted in regions where mapping does not satisfy the selection criteria. In order to save some time, if the processing of the whole reads file is too time-consuming, it is possible to create a sub-sample of the reads from a given sequencing run (as long as the total genome coverage remains high enough). To get homogeneously generated datasets, the {Sequencing_simulator} script creates artificial sets of reads from published sequence data. These artificial reads can then be used for assembly in BN. No sequencing error rate has been introduced in this simple tool. The script accepts as parameters the size of the artificial reads and the total coverage wanted. It runs via BioNumerics because it uses BN interface to be more user friendly.

For some purposes, it may be useful to produce a *de novo* assembly. This was done for instance for the ICDS analysis, or for the identification of deletions. As mentioned before, performing a *de novo* assembly can be a bit more tricky depending on the aim of the analysis. As for the homology assembly, the use of BioNumerics can be fairly simple as the basic steps just imply to load the reads, perform the assembly with one of the algorithms available (Velvet or Ray) and then retrieve the assembled contigs, potentially after displaying some statistics showing for instance the number of contigs and the total assembled length. But most probably the results of a basic assembly with default parameters will not be satisfying. The key element for *de novo* assemblies resides in the choice of the k-mer length used, noted k. It is mandatory to use different values of k (accessible via the BN interface in the Power Assembler module) as well as test different sets of reads (generated via a Python script {Reads_sampler} that sample the original data file to create some sub-samples).

In both types of assemblies, an important parameter that impacts the total percentage of the genome reconstructed as well as the quality of the mutations determined in the following steps is the minimum coverage expected to keep a region in the final assembly. It can be set to a small value (e.g. 1 or 2) if the goal is to reconstruct as many regions as possible, even if they are poorly covered or to a higher value (e.g. higher than 10) if one wants to insure a high-level of confidence for each base of the final sequence.

SNPs determination

For this work, the SNPs were determined by aligning the sequences obtained after the assembly by homology process. All sequences assembled by homology to a reference genome are perfectly collinear due to the way the assembly is performed. They will include "Ns" in regions absent or insufficiently covered from the target sequence, but present in the reference. The script {Snps_determination} used within BioNumerics will extract SNPs from the selected sequences stored in the BN database. It is based on the assumption that all sequences are collinear (*i.e.* result from an homology assembly on the same reference genome). Each position on the sequences is compared and SNPs are retained only if they are informative in all entries being compared (i.e. if the position does not correspond to a N in one of the compared sequences). The output from this step is a tab delimited text file, with SNPs coded as 1, 2, 3 or 4 (1 is A, 2 is C, 3 is G, T is 4, i.e. alphabetic order). This file can be imported into BioNumerics as a character data file.

SNPs filtering

When investigating phylogeny using SNPs, it is necessary to remove SNPs which can result from intrachromosomal genetic transfers as well as sequencing artifacts. Intrachromosomal genetic exchanges by recombination will by definition occur between homologous regions, i.e. any kinds of repeated sequences (gene families, IS elements, tandem repeats). In addition, such repeated elements may interfere with the assembly process. Reads coming from these regions can be assigned to other

homologous positions on the genome, even if there is one or two mutations differentiating them. In that case, the determination of SNPs from homologous regions of the genome will be erroneous and therefore these families of genes have to be excluded from the final mutation list. This is true for all genes that can be present at multiple copies on the genome, like tranposases or insertion sequences (and PE and PPE proteins in the case of the *M. tuberculosis* complex). Finally, some artifacts may be introduced by the sequencing technology itself and result in some abnormal clustering of SNPs (see below, "Identification and filtering of SNPs clusters").

The script {Mutation_table_analysis} takes an exclusion list in a text format file listing all the genes or positions on the genome that have to be filtered out, and then parses the file containing all the mutations to filter the unwanted ones. The exclusion list can include annotation key words of the reference strain (for instance to identify automatically all the positions on the genome corresponding to a given gene family) (Figure A1). The output file can then be imported into BN as a character dataset (i.e. displaying a matrix with strains on one side and the position of the SNPs on the other, and filled with the base for the considered strain at the specified position).

```
1  Exclusion list CDC1551 AC NC_002755.2 GI:50953765:
2
3  Families:
4  phiRv2 prophage protein
5  phiRV1 phage related protein
6  phiRv1 phage protein
7  phiRv2 prophage protease
8  transposase IS6110
9  phiRV1 phage protein
10 IS like-2 transposase
11 transposase
12 phiRv2 prophage integrase
13 integrase
14 phiRv1 integrase
15 IS1533 transposase
16 Bionumerics predicted
17 PE family protein
18 PPE family protein
19 PGRS family protein
20
21 Genes:
22 MT0695
23
24 Intergenic regions: #]first gene-last gene[, not included
25 MT0003-MT0004
26
27 Miscellaneous coordinates: #start   size (separator: tabulation)
28 3113900 3118150     #DR-locus
29 79486   79554
30 154283  154410
31 424072  424305
```

Figure A1: Example of the syntax of an exclusion list for the filtering script.

Genome annotation

In what has been described so far, the assembled genomes have not been annotated, the only information about annotation that has been used (for instance during the filtering process) has come from the annotation of the reference genome. It may be necessary to annotate the studied genomes, for instance to determine the type of the mutations (intergenic, synonymous or non-synonymous). In BN the Annotation module enables to determine the annotation of a sequence from the comparison with an annotated reference strain, by "transferring" the annotations based on an evaluation of synteny and homology. This will not annotate a gene that is not present in the reference annotation, but it is possible to input several annotated sequences in order to cover more of the genetic variety of a species pan-genome.

Genome annotation can be performed in BioNumerics using the "Annotation" module and following the instruction shown in the BN manual. This tool enables to "transfer" the annotations of a given number of reference genomes to a target sequence. As with several other aspects of working with NGS-produced sequences one problem is the time needed to annotate each sequence if the dataset considered is large. It can be done manually but it is a time-consuming task. In the case of SNP type determination, it is possible to speed up the process in order to know if a mutation is synonymous or not. This method relies on the creation (via a dedicated function in the script {Djib_update_generation}) of a "synthetic" genome sequence containing all the mutations previously identified and gathered in a global file. This sequence is stored in the BN database and annotated via the Annotation module, providing in a single step the mutation type of all the SNPs' dataset.

Construction of phylogenetic trees

The construction of trees is performed in BN via the "Comparison" module, which gives access to several different types of phylogenetic classifications using the most commonly accepted algorithms such as UPGMA, neighbor-joining (NJ), parsimony, etc. For our analyses we focused mainly on using Minimum Spanning Trees (MSTs) in order to display the relationships between the clonal strains, with

hypothetical nodes corresponding to the common ancestors of the strains present in the subtree considered. BN provides tools to assess the statistical relevance of a given classification. For SNPs one indication that was of great use was the total length of the branches in the MST, which differs from the total number of SNPs because of homoplasic mutations. This evaluation of homoplasy was a good indicator of the validity of the obtained tree, as the MTBC is considered to be clonal.

Identification and filtering of SNPs clusters

In species that are not clonal, genomes are susceptible to horizontal gene transfer. However these transfers alter the phylogenetic signal that would have been expected from a purely vertical evolution, and therefore they need to be identified, in order to generate accurate phylogenetic trees, based on the regions of the genome that were not affected by the transfers. This identification has been the subject of different works, and the methodology used in this work was the one proposed by Croucher *et al.* (Croucher *et al.*, 2011). This method evaluates the statistical significance of clusters of mutations in a given branch of a phylogenetic tree, with a null hypothesis that supposes a polynomial distribution of mutations in the absence of HGT. This method was transcribed into a Python script {Enhanced_tree_SNPs_clusters_analysis}, in order to be applied to datasets stored in BN and MSTs generated with these datasets. Given the evolution of the phylogenetic tree after filtering, it may be necessary to run this script several times until the structure of the tree reaches its final state.

This script takes many different parameters in order to perform the identification of segments of exogenous origin. The two major parameters are the size of the genome studied (in order to estimate the percentage of mutations on a per base basis), and the maximal size allowed for the window used during the cluster search. This script needs to be run on a given Minimum Spanning Tree. For each branch of the tree, it analyses the SNPs distribution and compares it with a polynomial distribution (the null hypothesis for the statistical test). The statistical significance between both hypotheses is evaluated, leading to the identification of exogenous clusters of

mutations that are listed and can be masked in the corresponding BN character dataset. The script includes ways to display the results of this analysis, for instance in order to identify the branches were HGT events have been found.

This method can also be used directly to filter the "raw" SNPs in a clonal organism: if there are no transfer, all the aggregated SNPs will correspond to noise in the phylogenetic signal and can therefore be masked. In that case, the identification of clusters is used as a filter to remove artificially generated clusters (technical artifacts due to the sequencing process, or the assembly and alignment of the sequences).

Miscellaneous

It is possible to identify the deletions between a reference genome and an assembled sequence by looking at the regions that have not been reconstructed during the homology assembly. In order to determine the position of these regions the script {Deletion_analysis} was written. This script produces a list of all the segments of Ns on the assembled sequence bigger than a given length that can be specified as a parameter. There are also auxiliary functions that enable to compare the deletions for different strains in order to identify shared ones. One peculiarity about this comparison is linked with the boundaries of these regions that have to be considered with a certain margin when comparing different strains in order to take into account the border effect caused by the mapping of the reads in conjunction with the minimal genome coverage for the assembly.

Hardware used for the present project:

The hardware equipment was a Getek machine (Courtaboeuf, France) with 24 Go of DDR3 RAM, a 240 Go OCZ Vertex 3 SSD drive, Core I7® extreme edition "Gulftown ", six cores 3.46 GHz and 2To internal hard disk for data storage, complemented by external storage disks with USB3 connection (2 external disks of 2To were used in the course of this work). The computer was running under Windows 7

Professional edition. The year 2011 price was 3000 € and it proved sufficient for the project requirements.

Software:

BioNumerics, software developed by Applied Maths NV Sint-Martens-Latem, Belgium.

Python programming language created by Guido van Rossum v.2.7 (running within BioNumerics) and v.3.2.1, for running scripts independently of BioNumerics.

Mauve, alignment software (Darling *et al.*, 2010) (http://gel.ahabs.wisc.edu/mauve/)

B) Details of other publications:

During my thesis I had the opportunity to work on several other projects besides the study of the origins of the *Mycobacterium tuberculosis* complex. These projects corresponded to the interests of the team in understanding the emergence and evolution of pathogenic bacteria, particularly in the light of the improved data accessible because of the development of high-throughput sequencing technologies. The methodologies developed during the study of the MTBC are easily transposable to the study of NGS data from other organisms.

The first of these pathogens is *Yersinia pseudotuberculosis*, the environmental bacteria from which the dangerous pathogen *Yersinia pestis* is thought to have evolved. *Yersinia pestis* had been particularly studied in the lab with the characterization of its CRISPR locus, leading to the study of its link with *Yersinia pseudotuberculosis*. The following publications focused on the *de novo* assembly of NGS sequencing data, leading to several publications in Genome Annoucement. The two following articles were the results of a collaborative work with Russian researchers on peculiar strains of *Yersinia pseudotuberculosis* collected in Russia. The third one results from the assembly of a strain isolated in a French hospital.

Article n°3:

Draft Genome Sequences of Two *Yersinia pseudotuberculosis* ST43 (O:1b) Strains, B-7194 and B-7195. Blouin Y, Platonov ME, Pourcel C, Evseeva VV, Afanas'ev MV, Balakhonov SV, Anisimov AP, Vergnaud G. **Genome Announc**. 2013 Jul 18;1(4).

Link to the article: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3715673/

Article n°4:

Draft Genome Sequences of Five *Yersinia pseudotuberculosis* ST19 Strains and One Strain Variant. Platonov ME, Blouin Y, Evseeva VV, Afanas'ev MV, Pourcel C, Balakhonov SV, Vergnaud G, Anisimov AP. **Genome Announc**. 2013 Apr 11;1(2):e0012213.

Link to the article: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624682/

Article n°5:

Yersinia pseudotuberculosis ST42 (O:1) Strain Misidentified as *Yersinia pestis* by Mass Spectrometry Analysis. Gérôme P, Le Flèche P, Blouin Y, Scholz HC, Thibault FM, Raynaud F, Vergnaud G, Pourcel C. **Genome Announc**. 2014, Jun 12;2(3).

The pair *Bacillus cereus* / *Bacillus anthracis* is another association of an environmental bacterium and a closely related virulent pathogen studied by our team. I had the opportunity to take part in a collaboration with the ANSES ("Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail") in the "Franthracis" project. The methodologies developed during my thesis were applied to the study of the phylogeny of French strains of *B. anthracis* to characterize the genetic diversity encountered in our country, where agriculture is still an important economic sector and breeding is practiced in many regions. Therefore the surveillance and management of anthrax outbreaks is crucial. This article was aimed at the determination of the population structure of the French strains from recent outbreaks, for instance linked with the recurrence of the events as well as with their localization. 122 strains collected over a 60-years period were sequenced at draft level, leading to the identification of 1581 SNPs that were used to construct the phylogenetic tree of these strains. Eventually a reduced set of SNPs was selected to be proposed as a highly discriminatory assay for French strains, for veterinary purposes.

Article n°6:

High-throughput sequencing of *Bacillus anthracis* in France: investigating genome diversity and population structure using whole-genome SNP discovery. Girault G, Blouin Y, Vergnaud G, Derzelle S. **BMC Genomics**. 2014 Apr 16;15(1):288.

Link to the article: http://www.biomedcentral.com.gate1.inist.fr/1471-2164/15/288

Due to the versatility of bio-informatics for analysis purposes I had the opportunity during my thesis to participate to another project that implied the analysis of sequences produced by NGS technologies. The following publications are linked with my participation to the genetic analyses of the bacteriophages studied in our lab as part of our efforts to understand the forces that shape bacterial populations. Many phages directed against *Pseudomonas aeruginosa* were collected in Abidjan (Côte d'Ivoire) and characterized by Christiane Essoh during her Ph.D, and I had the opportunity to assemble the genomes of these phages that were submitted to NGS sequencing. The *de novo* assemblies produced complete genomes that were then submitted to an extensive genetic analysis.

The first article considers the effects of several bacteriophages on a collection of *P. aeruginosa* strains from cystic fibrosis patients. Some of these phages have been submitted to whole-genome sequencing in order to get their full genomic sequence. The genera of the phages have been determined by electron microscopy and partial sequencing, showing a large diversity. Some bacterial strains from the collection were not lysed by any phage. The effect of the CRISPR system was investigated but it did not appear as a major defense mechanism in these resistant strains.

Article n°7:

The susceptibility of *Pseudomonas aeruginosa* strains from cystic fibrosis patients to bacteriophages. Essoh C, Blouin Y, Loukou G, Cablanmian A, Lathro S, Kutter E, Thien HV, Vergnaud G, Pourcel C. **PLoS One**. 2013 Apr 24;8(4):e60575.

Link to the article: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3634792/

The second article is centered on the detailed study of one particular bacteriophage from the "Abidjan collection". This temperate phage, Ab31, was characterized in conjunction with *P. aeruginosa* strains that had become resistant against it. The induced phenotypes are described precisely, and this phage has been submitted to whole-genome sequencing using high-throughput technologies. Its genomic sequence reveals that it is a chimeric phage with structural genes coming from a phage resembling the lytic phage AF of *Pseudomonas putida* whereas the regulatory portion of the genome is closely related to the temperate phage PAJU2, a phage infecting *P. aeruginosa*. The genomic analysis of some of the resistant strains *via* NGS revealed that the phage was inserted in stable lysogenes, and the insertion site was determined.

Article n°8:

A novel *Pseudomonas aeruginosa* Bacteriophage, Ab31, a Chimera Formed from Temperate Phage PAJU2 and P. putida Lytic Phage AF: Characteristics and Mechanism of Bacterial Resistance. Latino L, Essoh C, Blouin Y, Vu Thien H, Pourcel C. **PLoS One**

Link to the article: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3974807/

C) List of scripts written for the lab's projects:

Script name	Description
Alternative_mutations_analysis	Determines the SNPs between several selected strains by base to base comparison. There is no "alignment", the sequences are supposed to be collinear.
Annotation_update	Adds the annotations contained in one file to a given reference file (.gbk). The output is stored in a third file.
Automatic_groups_assignation	Clusters the entries from a given tree into groups according to a threshold chosen by the user. The tree is supposed to be stored in a BN comparison.
Automatic_read_extractor	Extracts reads from a fastq file according to the headers present in a temporary list.
BAPS_formating	Modifies a "classical" SNPs file for import into BAPS.
Basic_sequencing_simulator	Creates a set of reads from a genome in a fasta file. No implementation of sequencing errors.
Blast_command_window	Generates a window to input the files for a local BLAST (GUI for command line).
Blast_parser	Parses a BLAST text output in order to identify some particular information.
BN_to_Artemis_genome_conversion	Converts a BN .gbk file (in fact a file in .embl format) to a format that can be imported into Artemis (at least for the sequence part).
Branch_analysis	Analyses the SNPs contained in a reference file compared to the list of mutations belonging to a certain branch (construction of a projection on this branch).
Branch_analysis_utilities	Several functions dedicated to the analysis of snps belonging to a particular branch of a Minimum Spanning Tree generated with BN.
Can_data_generation	A collection of many different functions that were used during the study of the "*M. canettii*" dataset during the second semester of 2012.
CloneA_data_generation	Another collection of functions implied in the generation of the datasets and figures of the clone A article.
Clusters_analysis_H37Rv_limited	Script dedicated to the analysis of SNPs clusters by comparison to H37Rv, based on the use of a similarity matrix calculation.
Codon_usage_analysis	Analyses the usage frequencies of codons in a given annotated sequence stored in BN.
Comas2009_snps_analysis	Script used to compared SNPs with MLST results from Comas *et al.* 2009.
Contigs_extractor_multifasta	Extracts contigs stored in a BN sequence to a given multifasta file.

Contigs_extractor_multifiles	Extracts contigs stored in a BN sequence to a given folder, in separate fasta files (1 contig per file).
Contigs_filter	Generates statistics on contigs stored in a given BN sequence, and enables some size filtering on the contigs.
Contigs_fusion	Modifies a given multifasta file containing contigs in order to construct a single fasta sequence without any separations between contigs.
Contigs_stats	Generates statistics on contigs stored in a given BN sequence.
Create_reads_from_contigs	Enables to generate NGS reads from a multifasta contigs file.
Custom_selection	Enables to select entries in a BN database according to a given regular expression pattern.
Deletion_analysis	Compares sequences stored in BN and generated by homology assembly to determine the region of deletions between some strains.
Deletion_analysis_variant	Variant of the script "Deletion_analysis", based on a slightly different algorithm.
Djib_supp_tables_generation	A collection of many different functions that were used during the study of the Djibouti dataset during the second semester of 2012, and for the generation of the dataset and figures of the Djibouti article.
Dlg_library	A library of simple dialog boxes for BN interface.
DR_finder	Searches for DR sequences in reads in order to identify a majority of the reads covering the DR locus in a given set of sequencing reads.
EAI_snps_filtering	Filtering of EAI SNPs file.
Enhanced_tree_SNPs_clusters_analysis	Multi-purpose script dedicated to the analysis of SNPs clusters in MSTs generated with BN. Implements the algorithm described by Croucher *et al.* For the statistic identification of SNPs clusters based on a given tree.
Entry_duplication	Duplicates a given entry containing an annotated sequence in a BN database (annotations conserved).
Exclusion_list_generator	Formats the exclusion list used in the analysis of the Djibouti strains.
Exclusion_list_reader	Loads the exclusion list in order to use it to filter the SNPs for some Djibouti analyses.
Export_database_fields	Exports the information fields of a given BN database in a text file that can uploaded using "Import_db_fields".
Export_fasta_multifiles	Exports the sequences selected in a BN database as mulitple fasta files (one per sequence).
Export_genbank_multifiles	Exports the annotated sequences selected in a BN database as mulitple genbank files (one per sequence; the actual output is in .embl format).

Export_multifasta	Exports the sequences selected in a BN database as a single multifasta file.
Fastq_to_fasta	Converts a reads file in fastq format to a multifasta file.
Fields_insertion	Inserts a given information fields list contained in a text file into a BN database.
Fields_suppression	Deletes the selected information fields from a given BN database.
File_reading	Updates an annotated sequence with information contained in a tab-structured file.
File_transposition	Transposes a tab-delimited file as Excel could do (used for large files that cannot be opened in Excel).
Finishing_scan	Scans a newly assembled genome to identify abnormal bases (i.e. not A,C,G,T) and look into a read file to find reads that enable to correct these anomalies.
Gauss_filter	Filters the paired-end associated with the reads containing a given pattern, to suppress the Gaussian component of the signal when looking at the coverage of a genome.
GC_content_plot	Enables to plot the GC content of a genome, given a certain sliding window size.
GC_reads_analysis	Determines the GC content of a certain number of reads.
Gene_homoplasy_analysis	Analyzes the homoplasy for the Djibouty dataset by comparing with some "*M. canettii*" strains and other more distant mycobacteria (*M. avium*, *M. marinum*, etc).
Genome_filtering	Enables to delete some genomic regions from selected genomes; the resulting sequence is stored in a new BN experiment.
Global_reads_sampling	Cuts an input reads file in a given number of sub-samples stored in different files.
Graph_drawer	Displays segments on a graph.
HGT_clusters_annotation	Updates the annotation of an annotated sequence in a BN database with data from a flat text file.
Hgt_filter	Filters the SNPs contained in a file to identify HGT events.
Holes_filling	Uses a reads file in order to fill a gap in a de novo assembly of a given genome. Both boundaries are needed as a parameter.
Homoplasy	Analyzes a SNPs list in relation to a reference genome stored in a BN database in order to identify SNPs in genes.
Homoplasy_analysis_utilities	Collection of three functions useful for the analysis of homoplasic SNPs.

Hybrid_icds	Used to create an hybrid sequence during the study of the ICDSs.
Icds_analysis	Used during the early analysis of the ICDSs in the MTBC.
Icds_extraction	Extracts the ICDS subregion from a given genome stored as a sequence experiment in a BN database.
Icds_finder	Predicts potential ICDSs based on the position of the CDSs contained in an annotated file (Genbank format).
Import_database_fields	Imports the information fields of a given BN database from a text file that has been created using "Export_db_fields".
Import_MLST	Imports MLST data from a specific format into BioNumerics.
Import_SNPs	Used to import big SNPs dataset in a BN character experiment.
Indels_table_analysis	Analyzes a file containing indels positions according to an exclusion list as used for the Djibouti analyses.
Infection	Simulates the co-evolution of phages and bacteria in an environment, with a set of simple parameters.
Insertion_site_search	Enables to draw a graph of the GC content of a genome in order to look for the insertion site of a phage in a lysogene.
Last_line_suppression	Deletes the last line of a file. Useful in some cases: reads sampling, reads filtering with BN, etc.
Mult_align_snps_generator	Generates a concatenated nucleotide sequence from a file containing SNPs, and stores the results in a multifasta file.
Mutations_occurrences_analysis	Analyzes the mutations on the branch of a MST stored as a BN comparison in order to be able to draw diagrams of the mutations types for each branch of the tree. Used during the Djibouti analyses.
Mutation_table_analysis	Modifies the mutations list obtained from the Chromosome Comparison of BN in order to take into account the Ns present in homology assemblies and to filter the results according to an exclusion list written for a reference genome. Used during the early stages of the Djibouti analyses with H37Rv as a reference.
NGS_250bp_reads_division	Divides a 250bp-reads dataset from a given fastq file in 125bp reads stored in a single fastq file.
NGS_250bp_reads_treatment	Extracts single reads from any reads-containing file. The quality is not conserved. Developed for phage assembly using 250bp reads.
Ns_suppression	Suppresses Ns in a given BN sequence. Use with caution, as Ns can be meaningful.

Pattern_finder	Searches for a given sequence string ("pattern", can be a regular expression) in all the selected sequences in a BN database.
Phages_site_update	Updates a Django web site with the data contained in a BN database. Development version.
Phage_annotation	Modifies the annotation of a BN annotated sequence in order to follow "rules" used for the annotation of some of the lab's phages genomes.
Phage_cosmetics	Similar to phage annotation, but for "cosmetic" aspects needed for EBI submission.
Phage_insertion_search	Parses the output files of two specific BLAST results in order to identify hybrid reads characteristic of the insertion of a phage in a bacterial genome.
Reads_analysis	Searches in a reads file for a given DNA pattern (treated as a regular expression pattern).
Reads_analysis_BN_free	Same function as "Reads_analysis", but without the need of using BioNumerics (but also without interface).
Read_extractor	Extracts a read from a fatsq file according to its header.
Reads_sampler	Takes a subsample of a global fastq reads file. This subsample is stored in a new fastq file.
Sequencing_simulator	Generates simulated reads (without errors) from a sequence stored in a BN database.
Similarity_matrix_generation	Generates a similarity matrix by pair-wise comparison of strains based on a file containing SNPs. Iterates on this matrix with the statistical test from Croucher *et al.* In order to filter the clusters on each "branch" (between two given strains).
Similarity_matrix_generation_BN_free	Same as "Similarity_matrix_generation", but without the need to use BioNumerics. Implements slightly more functions to improve the last steps of the analysis.
Snps_cluster_analysis	Very basic cluster prediction from the analysis of a file containing SNPs positions.
Snps_cluster_filtering	On the same basis as "Snps_cluster_analysis" this script enables to filter clustered SNPs for a file containing the SNPs positions.
Snps_determination	Determines SNPs from base to base comparison of homology assemblies stored as sequences in BN.
Snps_histogram	Returns a file containing the data to draw an histogram with Excel of the frequencies of SNPs in a given window size.
Snps_mauve_analysis	Treats SNPs contained in a Mauve result file, and enables for instance the attribution of these SNPs to given CDSs as determined from a reference contained in a Genbank-formated file.

Spa_types	Enables the automated analysis of MLVA data sequences for *S. aureus*.
Spoligo_finder	Adaptation of the "Reads_analysis" principle in order to look for the DR locus in a given reads file. The reads collected with this script can be used to reconstruct the locus *de novo*.
temp	Collection of miscellaneous useful functions constructed when needed. Repository of functions during development.
Test_BN	Collection of several functions meant to test different BN functionalities and interactions between the Python scripts and the main software.
Transfer_analysis	Identifies the regions originating from an event of horizontal gene transfer in a given genome, with some basic descriptive statistical indications.
Tree_analysis	Some functions to analyze a MST stored in BioNumerics.
Tree_hgt_branch_filtering	Identifies and enables the filtering of HGT events from a BN MST.
Tree_homoplasy_evaluator	Analyzes the homoplasy on given BN-generated MST. Returns a list of homoplasic SNPs and the branches where they occur.
Unknown_phage_annotation	After CDSs prediction with BN this script performs a BLAST of all CDSs on the NCBI database and annotates the CDSs with the best BLAST result.
Utilities	Collection of several miscellaneous functions useful for some file manipulations or data handling. Repository for functions that have been needed at some point during certain steps of an analysis of genomic data.
VNTR_NGS_evaluation	Searches for VNTR reads in a reads file. Same concept as the basic reads analysis, takes as input a file containing the consensus sequence for each repeat.
Ypes_sra_mod	Collection of functions used during the analysis of *Y. pseudotuberculosis* strains from Russia. Some functions enable to check that different genomes assembled on the same reference are collinear, and to correct it if they are not.